V PSYCHOLOGICAL REVIEW PU S WHOLE No. 204
1934

chological M ographs

EDITED BY

HERBERT S. LANGFELD, PRINC UNIVERSITY

ACTICE AND V ABILITY

A Study in Psycholo Method

BY

ANNE ANAST

INSTRUCTOR IN PSYCHOLOGY, LLEGE

UBLISHED FOR THE AMERICAN PSYC ASSOCIATION BY
PSYCHOLOGICAL REV MPANY
PRINCETON,
AND ALBANY,

Publications of the American Psychological Association

PSYCHOLOGICAL REVIEW

containing original contributions only, appears bi-monthly, January, March, May, July, September, and November, the six numbers comprising a volume of about 540 pages.

PSYCHOLOGICAL BULLETIN

containing critical reviews of books and articles, psychological news and notes, university notices, and announcements, appears monthly (10 numbers), the annual volume comprising about 720 pages. Special issues of the BULLETIN consist of general reviews of recent work in some department of psychology.

JOURNAL OF EXPERIMENTAL PSYCHOLOGY

containing original contributions of an experimental character, appears bi-monthly, February, April, June, August, October, and December, the six numbers comprising a volume of about 900 pages.

PSYCHOLOGICAL INDEX

is a compendious bibliography of books, monographs, and articles upon psychological and cognate topics that have appeared during the year. The INDEX is issued annually in June, and may be subscribed for in connection with the periodicals above, or purchased separately.

PSYCHOLOGICAL ABSTRACTS

appears monthly, the twelve numbers and an index supplement making a volume of about 700 pages. The journal is devoted to the publication of non-critical abstracts of the world's literature in psychology and closely related subjects.

PSYCHOLOGICAL MONOGRAPHS

consist of longer researches or treatises or collections of laboratory studies which it is important to publish promptly and as units. The price of single numbers varies according to their size. The MONOGRAPHS appear at irregular intervals and are gathered into volumes of about 500 pages.

JOURNAL OF ABNORMAL AND SOCIAL PSYCHOLOGY

appears quarterly, June, September, December, March, the four numbers comprising a volume of 448 pages. The journal contains original contributions in the field of abnormal and social psychology, reviews, notes and news.

ANNUAL SUBSCRIPTION RATES

Review: $5.50 (Foreign, $5.75). **Index:** $4.00 per volume.
Journal: $7.00 (Foreign, $7.25). **Monographs:** $6.00 per volume (Foreign, $6.30).
Bulletin: $6.00 (Foreign, $6.25). **Abstracts:** $6.00 (Foreign, $6.25).
Abnormal and Social: $5.00 (Foreign, $5.25). Single copies $1.50.
Current numbers: Journal, $1.25; Review, $1.00; Abstracts, 75c; Bulletin, 60c.

COMBINATION RATES

Review and Bulletin: $10.00 (Foreign, $10.50).
Review and J. Exp.: $11.00 (Foreign, $11.50).
Bulletin and J. Exp.: $12.00 (Foreign, $12.50).
Review, Bulletin, and J. Exp.: $16.00 (Foreign, $16.75).
Review, Bulletin, J. Exp., and Index: $19.00 (Foreign, $19.75).

Subscriptions, orders, and business communications should be sent to the

PSYCHOLOGICAL REVIEW COMPANY
PRINCETON, N. J.

VOL. XLV
No. 5

PSYCHOLOGICAL REVIEW PUBLICATIONS

WHOLE NO. 204
1934

Psychological Monographs

EDITED BY
HERBERT S. LANGFELD, PRINCETON UNIVERSITY

PRACTICE AND VARIABILITY

A Study in Psychological Method

BY

ANNE ANASTASI

INSTRUCTOR IN PSYCHOLOGY, BARNARD COLLEGE

PUBLISHED FOR THE AMERICAN PSYCHOLOGICAL ASSOCIATION BY
PSYCHOLOGICAL REVIEW COMPANY
PRINCETON, N. J.
AND ALBANY, N. Y.

TABLE OF CONTENTS

CHAPTER I

Present Status of the Problem *

With the development and widespread use of standardized tests to measure individual differences, problems arise as to just what these measures tell us about the individual. The most obvious interpretation of test results is that they show the status of the individual at the time in the activity tested, reflecting all the past experience and training of the individual to date, as well as any momentary conditions which might raise or lower efficiency. The differential psychologist, however, would like to know more than this about his subjects; from the sample of behavior measured by his test, he hopes to predict, to some extent, future performance of the individual. Data on test reliability throw some light on the amount of fluctuation to be expected from momentary conditions of the subject. The question of how far the subject's performance is determined by differences in amount of specific training in the same or similar activities in the past, and how far it is determined by individual differences in the *response* or susceptibility to such training, is still very difficult to answer. Test makers have usually made vague attempts to settle the question for their particular test by such devices as including only material which is fairly common in the experience of "most people", and then by inserting cautious warnings to use the test only on groups "similar" to those upon which the test was standardized. Such procedures may succeed in evading the difficulty to a certain extent and may give usable test results for most practical purposes; but obviously a more thorough analysis of the question is needed before test results can be taken very seriously.

It is probably this question, more than any other factor, which

* This study was made possible through a grant from the Council for Research in the Social Sciences of Columbia University.

stimulated early research on the effect of practice on individual differences. Thorndike (22), in one of the earliest investigations of this problem in 1908, expresses the significance of such work as follows: " Experiments in practice offer evidence concerning the relative importance of original nature and training in determining achievement. In so far as the differences amongst individuals in the ability at the start of the experiment are due to differences in training, they should be reduced by further training given in equal measures to all the individuals. If, on the contrary, in spite of equal training the differences amongst individuals remain as large as ever, they are to be attributed to differences in original capacity." In his *Educational Psychology* (24, Vol. III, pp. 304–5), he makes practically the same statement before reporting the available data on practice and individual differences to date. Wells (29) begins his article on " The Relation of Practice to Individual Differences ", by stating a set of " postulates ", *viz:* (1) " the nearer a subject is to the end of his practice curve, the less the prospect of improvement, and the farther he is from the end, the greater the prospect of improvement. (2) If high initial efficiency is the product of *greater practice,* expectation of further practice improvement is small. (3) If high initial efficiency is the product of a greater ability to profit by a given amount of practice, then the expectation of improvement under special practice is great." Hollingworth (4), in the introduction to an article entitled " Individual Differences Before, During and After Practice ", states: " In the direct application of mental tests, it has too often been assumed that the actual performance of an individual in one or a dozen trials at a given task is, in some way or other, significant of that individual's final capacity for such work. It is true that several investigators . . . have studied the effects of practice on individual differences. . . . Such studies have produced suggestive results, although they have been based, for the most part, on records of only a few subjects, or on a relatively small number of practice trials."

These statements are typical of the attitude of many of the earlier workers on the problem, and indicate that their work was motivated largely by an interest in whether individual differences in test performance result from different amounts of specific past training, or from differences in the susceptibility to such training. Most writers, furthermore, tie this question up with that of heredity and environment, differences in susceptibility to training being identified with hereditary causes of individual differences, and differences in amount of specific training with environmental causes. The question of hereditary versus environmental causes of individual differences need not concern us here. It cannot be settled by the sort of data obtained in a study on the effect of practice upon individual differences. What we mean by " susceptibility to training " might very well be interpreted, if the reader may so desire, as the influence of more general environmental differences other than specific training in the function tested (or closely similar functions), which made the individual more susceptible to training of this sort in the past, and will also make him so in the future. The problem can be regarded entirely from the standpoint of the feasibility of prediction of future performance from present performance, without any entanglements in the moot question of heredity versus environment.

The literature on practice and individual differences has been so extensively reviewed by Kincaid (7), Peterson and Barlow (12), and Reed (14), that to repeat here a statement of results obtained by former investigators would be highly redundant. Each of the above reviewers analyzes the data published to date with special reference to the type of measures used to express variability. Kincaid reports, for each of the studies surveyed, seven measures * of absolute as well as relative variability. Peterson and Barlow give only the coefficient of relative variability,* (V), for each of the studies quoted. Reed gives, besides V, the ratio of the average of the three highest and the three lowest scores at the beginning and end of practice, and the correlation between initial trial score and per cent gain.* In addi-

* Some of these measures are not reported for all of the studies quoted, since the information in the published data was inadequate for their computation by the reviewers.

tion to this, each of the three reviews contains a more detailed report of new studies carried out by the reviewers.

An analysis of the literature on this problem reveals, besides a tangled mass of conflicting conclusions and apparently conflicting results, a set of clear-cut methodological controversies. The most outstanding of these centre about questions of how to express the results obtained. Most investigators now agree that the chief cause of the inconsistent results reported is the use of different methods of expressing the data. When the data are all couched in the same statistical form, the inconsistencies disappear to a remarkable extent, remarkable in view of the widely different subjects used, from school children to graduate students, and the variety of tasks practiced, from ball-tossing to mental multiplication and from cancelling A's to learning history. With regard to this last factor, the nature of the task practiced, Peterson (12) has suggested a theory that in the "simpler" processes, subjects tend to converge, or become more alike, with practice, and in the more "complex" they tend to diverge, but his own subsequent results do not fulfill this expectation, as he himself points out. Chapman (1) had suggested essentially the same hypothesis in 1914, offering as evidence the correlations between initial score and gain, which were all low but tended to be positive more often for complex than for simple tests. His data, however, are rather meagre, especially since none of the correlations was reliable when evaluated in terms of its P.E.

Since the problem of practice and variability is at present in such a controversial state, and since the present study was undertaken largely from a methodological standpoint, a survey of what seem to be the fundamental methodological issues in the problem should prove illuminating. There are at least five main issues which recur most frequently in the experimental as well as in the purely argumentative literature on practice and variability.

1. *Equal Amounts of Practice.*

By equal amounts of practice, most investigators mean equal time spent in practice. Thus, if all the subjects were given 100 trials of 10 minutes each, they would be said to have had the same amount of practice in the course of the experiment. It has been pointed out repeatedly by Peterson (10, 12), Reed (14), and others that this does not mean that the subjects have all done an equal amount of *work* on the task practiced, since the faster subjects do more actual work in the given time than the slower subjects. Does this mean that the subjects who are faster at the start receive more practice because of the nature of the experimental set-up? Obviously, the experimenter should make clear at the outset what he means by equal amounts of practice.

Results obtained by the "work constant" method are not strictly comparable with those obtained by the "time constant" method. The former tends to give less practice to the better subjects and relatively more to the poorer ones than does the latter. For this reason, the "work constant" method would tend to emphasize *convergence* and the "time constant" method *divergence* of individuals with practice. As long as the experimenter defines clearly what he means by equal amounts of practice in each case, either procedure seems theoretically justifiable. It might be suggested, however, that since the "work constant" method usually implies expressing performance in terms of *time scores,* and the "time constant" method in terms of *amount scores,* the latter should be preferred. The relative merits of these two scoring techniques will be considered in the following section.

2. *Time Versus Amount Scores.*

The fact that when the same set of data is expressed as time required per unit of work, or as amount of work done per unit of time, opposite conclusions can be drawn regarding the effect of practice on variability, has been stressed repeatedly in the literature and demonstrated anew by various writers. Whitley (30) gives a good analysis of the inconsistency found, illustrating her discussion with hypothetical examples on five subjects. She shows that with the given set of data, time scores may show a decrease in variability with practice, while amount scores show an increase. Wells (29) gives essentially the same theoretical analysis of the problem. Peterson (11), in criticizing Thurstone's conclusion that absolute variability measured by the S.D. increases with practice, based on the data from 165 men studying telegraphy, again points out that results from the *same subjects* may show convergence when expressed as amount scores, and divergence when converted into time scores.

Reed (13) demonstrates the same fact, using hypothetical illustrations very similar to those of Whitley. Chapman (2) carries the analysis further than had heretofore been done by showing *why* the inconsistency is found. With a cleverly devised hypothetical example, based on Reed's illustrations, he shows that when measuring performance from an arbitrary zero, as is done in nearly all present tests, time and amount scores yield inconsistent results; but this inconsistency disappears when we add into both sets of scores some constant representing the distance of the arbitrary zero from absolute zero. The inconsistency arises from the fact, which Reed had also pointed out, that in both time and amount scores we are dealing with fractions, not absolute quantities. The scores are actually amount of work *per unit time,* or time *per unit work.* In the former the unknown quantity which represents the distance from arbitrary zero to absolute zero occurs in the numerator, and therefore can be factored out as a constant; in the latter, it cannot be so factored out, and will therefore affect each score differently. Chapman concludes from this analysis that amount scores are theoretically sounder than time scores, when performance is measured from an arbitrary zero point.

3. *Absolute Versus Relative Measures of Variability.*

Reviewers have pointed out that when variability is expressed by some absolute measure such as Q, A.D., S.D., gross gains made by initially high and low individuals or groups, or correlation between initial standing and gross gain, then in most cases variability seems to *increase* with practice. When, on the other hand, relative measures of variability are employed, such as V, or some measure making use of relative or per cent gains, or ratios between the scores of initially high and low individuals, then variability seems to *decrease* with practice in the majority of cases. Kincaid (7) pointed out that this was the chief cause of the apparent discrepancies in the results of different investigators. The same fact was indicated by Peterson and Barlow (12) and by Reed (14), who, however, go farther and conclude that only *relative* measures are justifiable. The argument in support of relative measures is that, since the numerical size of scores changes during practice, it is as if the scores were expressed in different units in the different trials, and hence absolute measures are not directly comparable. Through chance alone, the argument runs, absolute variability will *increase* when the size of scores increases, and *decrease* when the scores decrease, hence such changes in absolute variability are of the nature of a statistical artifact.

The discrepancy in results obtained with absolute and relative measures of variability received attention in the earliest studies on practice and variability. Whitley (30) mentions it in her methodological analysis of the problem, and concludes that we should "prefer gross to percentile* measures of the ability in question" (p. 109). She illustrates her argument with hypothetical numerical examples. Wells (28) criticizes the use of percentage gains on the ground that it "runs counter to the fundamental conception of practice curves", referring by this to the phenomenon of negative acceleration. To improve from 600 to 900, he states, represents much more progress than to improve from 100 to 150, and yet the percentage gains in the two cases would be equal. Turning to more recent studies, we find Peterson repeatedly arguing for measures of *relative variability,* and employing them in all of his studies (cf., *e.g.,* 10, 11, 12).

Reed (13) also advocates the use of relative measures, particularly because of the inconsistency in measures computed from time and amount scores, which results when absolute measures of variability are used. The question of this inconsistency has already been taken up in the preceding section. Chapman (2) criticizes the use of relative measures, referring especially to Reed's article, on the grounds that the computation of such measures assumes that ability has been measured from an absolute

* Percentage.

zero. Reed (14) answers this criticism in his review as follows: " In reply, it may be stated that the ratio method, just like every other method, would increase in validity, if used on a scale having an absolute zero, but when applied to scores from an arbitrary scale, it is just as valid as other measures of this sort " (p. 14).

With the above statement we must take issue. It is a well-known fact that scores measured from an arbitrary zero can justifiably be used for many purposes, since they vary from scores measured from absolute zero by a *constant* amount. Hence, conclusions based on certain relationships among such scores will not necessarily be in error. The S.D., for example, will not differ when computed from scores measured from arbitrary or from absolute zero-points. The only measures that *will* yield misleading results when an arbitrary zero-point is used are *ratios* and *quotients.* Such measures will not differ from the measures on an absolute zero scale by a *constant quantity,* but each will be affected differently, even to the extent of a complete reversal of the relationships between them. This necessary limitation in the use of scores measured from arbitrary zero has been pointed out repeatedly (cf., *e.g.,* 26, p. 339). Since, however, it is of especial significance in the problem of practice and variability and has frequently been overlooked or misunderstood, it will not be amiss to illustrate it with the following examples.

Let us suppose that we are measuring performance on a scale on which the arbitrary zero is 20 points removed from absolute zero, when this distance is expressed in the same units as scores on the scale. The effect of using arbitrary zero upon the value of V computed from scores on this scale is shown in the four situations outlined in Table I. It will be noticed that the use of an arbitrary zero always makes V larger. The most significant fact, however, is that this increase in size of V is not constant, but in each of the examples given, the increase is greater the lower the mean. Since, when amount scores are used, the initial mean is lower than the final mean, the use of an arbitrary zero will increase the size of the initial V more than that of the final V in such cases. Such an error, then, is not only misleading because it affects different V's differently, but it actually loads the dice in favor of the interpretation that practice decreases individual differences.

It should be pointed out that the relationship between size of mean and increase in V resulting from the use of an arbitrary zero is not as simple and direct as the reader might suppose from a cursory examination of the four examples given. In cases where V is much larger after practice than before, the use of an arbitrary zero may increase final V more than it does initial V. If S.D.'s were constant and only the means varied, then the relationship would be very direct; the higher the mean the less would V be increased by the use of arbitrary zero. But since S.D. increases with the mean, both numerator and denominator vary, complicating the effect. In our examples, for instance, if the final mean and S.D. from arbitrary zero are 60 and 15, respectively, giving a V of 25, then V computed from absolute zero will be 18.75. The difference between 20 and 25, the initial and

TABLE I

THE EFFECT OF ARBITRARY ZERO UPON V

Case I: *V computed from arbitrary zero decreases with practice*

	Scores Measured from Arbitrary Zero		Scores Measured from Absolute Zero	
	Initial Trial	Final Trial	Initial Trial	Final Trial
Mean	50	80	70	100
S.D.	10	15	10	15
V	20	18.75	14.29	15

Case II: *V computed from arbitrary zero remains constant with practice*

	Scores Measured from Arbitrary Zero		Scores Measured from Absolute Zero	
	Initial Trial	Final Trial	Initial Trial	Final Trial
Mean	50	75	70	95
S.D.	10	15	10	15
V	20	20	14.29	15.79

Case III: *V computed from arbitrary zero increases with practice*

	Scores Measured from Arbitrary Zero		Scores Measured from Absolute Zero	
	Initial Trial	Final Trial	Initial Trial	Final Trial
Mean	50	70	70	90
S.D.	10	15	10	15
V	20	21.43	14.29	16.67

final V from arbitrary zero, is greater than that between 14.29 and 18.75, the corresponding V's computed from absolute zero. In this case, arbitrary zero has served to make the increase in variability with practice appear larger than it actually is. Thus a point could be found for any numerical set-up at which the difference between initial and final V's is the same, whether arbitrary or absolute zero is used, and beyond which the effect of using arbitrary zero is reversed. This point, however, is usually reached when V from arbitrary zero is considerably greater after practice than before, so that it would not enter into the results ordinarily found in experiments on practice.

So far our argument for absolute measures has been purely negative. Since statistical analysis has shown that relative measures rest upon invalid assumptions, the conclusion is that we have no choice but to use absolute measures. Something can, however, be said in favor of absolute measures on their own account. We cannot agree fully with Reed's statement that, "An S.D. has meaning only in relation to the average from which it is computed. . . . Its size is dependent on the size of the average and of the cases from which it is computed" (14, p. 19). This is true in so far as S.D.'s computed from scores in different units are not comparable; but *a change in size of scores does not in itself imply a change in units.* It is true that of two tests given to the same group, both of which yield a normal distribution and hence are suited in difficulty range to this group, the test with numerically higher scores will yield a higher S.D.

A very obvious example of this same point is the fact that the S.D. of a distribution of *time scores* will be 60 times as great when the scores are expressed as seconds as when the same scores are expressed as minutes. This situation, however, is not strictly parallel with that in which the *same* test is given either to different groups or to the same group under different conditions, such as before and after practice. Suppose we are comparing variability on two tests A and B, and that A has a possible maximum score of 25, B of 50 points. Then, by chance alone, scores could range from 0 to 25 on test A and from 0 to 50 on B. This would tend to make the absolute variability of B higher than that of A. If, now, we turn to the second type of case, we find a different situation. Suppose the highest score on a test C made by group I is 25, and the highest score made by group II on the same test is 50, it does not necessarily follow that group II will have a higher S.D. than group I. In fact, group II might have a lower S.D., if, for example, group II consisted of superior 6th grade children in a private school, and group I of unselected 3rd grade children in a public school.

Thorndike illustrates this same point in *Mental and Social Measurements* (23, p. 133). "In some cases," he writes, "the factors which make the central tendency larger seem to work to make the variability actually smaller. Thus, if from the same race living under the same conditions a group of tall men and a group of short men are picked (at random as far as variability is concerned) by picking men with very long fingers and men with very short fingers, the tall men show a gross variability that is *less* than that of the short men."

As a final example, let us consider two groups of men both with an S.D. for height of 8 inches, but with average heights of 64 and 72 inches in groups I and II respectively. Two such groups could no doubt be assembled without much difficulty. Would it not be a *reductio ad absurdum* to insist that group II is less variable in height, " actually ", than group I, and that the inches used in measuring height had in some mysterious fashion changed in value when we passed from one group to the other?

It seems to the writer that the measurement of a group before and after practice with a given test falls more nearly under the second of the two types of situations described above than under the first. Let us suppose that the highest score made by our group on trial I is 25, making the possible range from 0 to 25, and that on the last trial the highest score is 50. This does not mean that the possible range of scores on the last trial is from 0 to 50, since all subjects will have made some improvement. This seems a justifiable assumption, since we are presumably studying the effect that improvement through practice has on different subjects. Statistically, it would be quite possible for all the subjects to improve in such a manner that the entire range would shift up 25 points from the first to the last trial, without changing in size.

As for the objection that units do change in the course of practice and hence are not strictly comparable throughout, it would seem that the use of relative measures would tend to increase rather than decrease the error resulting from such a situation. Students of learning curves have frequently suggested that the improvement of one unit at the end of the practice period indicates greater progress than the improvement of one unit earlier in the course of practice. In other words, units become *larger* during the course of practice. The use of relative measures of variability, however, has the effect of giving each unit *less* weight after practice than before. Wells seems to have had something of this sort in mind in his criticism of relative measures of variability quoted earlier (p. 7).

4. *Relationship of Initial Standing to Improvement.*

Measures of the relationship between initial standing and improvement have frequently been used to express the effect of practice upon individual differences, either by themselves or in addition to measures of variability. Such measures have been criticized from three different angles.

(a) *Physiological Limit:* Most correlations obtained between initial scores and absolute gains have turned out to be negative. It has been argued repeatedly that this could not be otherwise, and that it results from the fact that as subjects approach their " physiological limit ", their chances of improvement become progressively less. The implication is that those who stand highest at first are nearer this limit and hence will improve less than those who start out lowest. Thus Stoddard (**18**, p. 480), in a general critique of the problem, states: " It is possible in such simple material as addition, cross-out tests, etc., that physiological limits are reached, or that the performance at least suffers through distinctly diminishing returns. Suggested possible limits are: visual-motor movements (eye-hand coördination) and capacity for effective attention in monotonous repetition."

The term " physiological limit " seems to have been used rather loosely to cover two somewhat different concepts. First, it is used in the sense of the level of final attainment of the subject in the given performance, the very best the subject can ever do with the task under the given experimental conditions. In this sense, the physiological limits of different subjects might be very far apart. The subject who starts highest might be just as far or even farther from his physiological limit on trial one, than the subject who starts lowest. Hence, the initially high subject should have as much chance of improvement as the initially low.

More frequently, the term physiological limit is used to denote the fact that in many tests, performance is determined more largely by speed of motor and sensory processes after practice than it is before practice. This point will, indeed, be reached sooner by the initially best subjects. For example, if the task were the mental multiplication of two-place by one-place numbers, individual differences in speed of writing would play a relatively small part in the initial trials. With practice, however, a point would soon be reached at which the subject "got the answer" as quickly as he could possibly write it down. Further improvement in speed of multiplying beyond this point would not show up in performance on this test. This point would certainly be reached sooner by the initially better subjects, and their chances of improvement would therefore be less than those of the poorer subjects.

This does not mean that the better subjects have reached their level of final attainment in multiplying sooner than the poorer subjects. It simply means that the test used is poorly chosen for a study of the effects of practice upon individual differences, since in the early stages of practice, performance on the test is determined chiefly by processes in which individual differences are large (*viz.*, speed of multiplying), and in the later stages, by processes in which individual differences are relatively small (*viz.*, speed of writing).

(b) *Relation to Group Variability:* Reed (14) has objected to the use of correlations between initial scores and gross gains on the grounds, first, that it may lead to conclusions opposite from those reached with correlations between initial scores and *relative* gains, and secondly, that such a correlation is not always a correct index of changes in group variability. In reply to the first criticism, we might say that the criticism can be directed equally well against the use of correlations between initial score and relative gain. We grant that the results obtained with the two correlations may be opposite, but in such a case we should rather throw out the measure based upon *relative* gains, for reasons already stated in section 3.

The second point merits further analysis. According to Reed, " Negative correlations between initial performance and gross gain . . . would be indications of *reduced variability,* but a positive correlation between initial performance and gross gain . . . may mean either an increase or a decrease in variability " (**14**, pp. 19–20). Reed argues thus: If a subject with a high initial score makes a larger absolute gain than a subject with low initial score, there will be a positive correlation between initial performance and gross gain. This, however, does not indicate convergence, since the gain made by the initially higher subject may have been *relatively* smaller than that made by the initially lower, in which case " the two will eventually come together " (**14**, p. 20).

What the above statement could mean is difficult to see. It certainly is not true that convergence of the raw scores is compatible with such a condition. *As long as* the absolute gains made by the higher subjects are larger than those made by the lower ones, irrespective of relative gains, the S.D. must increase and the curves plotted with the raw scores must diverge. Reed may have been referring to the fact that when the better subject gains *relatively* less than the poorer one, although at first the difference between them may increase, a point will be reached beyond which the difference decreases. At this point, however, the *absolute* gain made by the better subject will be *less* than that made by the poorer subject, hence the nature of the absolute gains is perfectly compatible with the resulting convergence. This situation may be more clearly understood by examining the example presented in Table II.

Two subjects, A and B, make initial scores of 50 and 20, respectively, and their rates of gain are 1/10 and 2/10, respectively, throughout the practice period. As long as the absolute gains of B are larger than those of A, the scores diverge, but in the fifth trial they begin to converge, A having gained only 6.66 from trials four to five, against B's 6.91. All this, however, does not invalidate conclusions based upon the correlation between initial score and gross gain. In the above example, the correlation between score on trial 1 and gross gain from trial 1 to trial 5 will be positive, and when comparing trials 1 and 5, we *do* find divergence, the difference between the two subjects increasing from 30 to 31.74. The convergence occurs from trials 4 to 5, and accordingly, if we correlate score on trial 4 with gross gain from trials 4 to 5, we will get a negative correlation. Similarly, trial 7 shows convergence when compared with trial 1, the difference between A and B being 28.87 as against 30 in trial 1. The gains made by A and B from trial 1 to trial 7 are 38.58 and 39.71, respectively; hence the correlation between initial score and gross gain would be negative, and again consistent with the measure of variability.

TABLE II

CONVERGENCE AND DIVERGENCE OF SCORES WITH PRACTICE

Trial	1	2	3	4	5	6	7
A	50	55	60.5	66.55	73.21	80.53	88.58
B	20	24	28.8	34.56	41.47	49.76	59.71
Diff. (A–B)	30	31	31.7	31.99	31.74	30.77	28.87

The upshot of this discussion seems to be that positive as well as negative correlations between initial score and gross gain are valid indices of divergence or convergence, respectively, *for the segments of the curves from which they are derived,* but that if the gains are measured over a narrow portion of the practice curve, then we cannot necessarily predict, from the correlation obtained, what is happening to the rest of the curve. In other words, such correlations are a perfectly valid index of changes in variability for the data upon which they have been computed.

The correlation between initial score and *relative* gain does not, in fact, tell us any more than that, although it might seem to from the above illustration. A negative correlation between initial scores and relative gains is indicative of ultimate convergence of the practice curves *only* if the rate of improvement remains constant, which seems very unlikely from our knowledge of practice curves. The use of gross gains, then, is fully as informative as the use of relative gains in computing the correlation between initial standing and improvement, and it has the advantage of being free from the assumption of absolute zero in the measures used.

(c) *Errors of Measurement:* It is now a well-known fact that errors of measurement serve to " attenuate " the coefficient of correlation between two sets of measures. This effect is usually to bring about a numerical reduction in the size of r, from what it would have been, had " true " scores been used. In correlations between initial scores and gains, however, the effect is more complex. Errors of measurement in this case may change a correlation from positive to negative and, up to a certain point, may *raise* the numerical value of a " true " negative correlation.

This effect was first pointed out by Thorndike (25), who illustrated it with a specific example, but offered no mathematical correction for it. Thomson (20) first reported a formula for correcting the correlation between initial score and gain for errors of measurement. In its final form (21) this formula involves only the reliability coefficients and sigmas of the initial and final trials, and the obtained correlation between initial and final trial. Thomson also shows graphically the relationship between true and obtained correlations, which brings out the fact that the relationship is not as direct as in the case of ordinary correlations. The complicating factor is the presence of the *same* chance error, as a positive quantity in the initial score of each subject and as a *negative* quantity in the *gain* made by the same subject. This produces a negative correlation between initial score and gain when the true correlation is actually zero or low positive.

Syrkin (19) calls attention to the fundamental importance of this attenuation effect in the problem of practice and variability. He states: " Dans plusieurs recherches, on a obtenu le résultat

$r_{\delta x} < 0$. Peut on conclure . . . que l'on était en présence du phénomène de convergence? Le but des pages qui vont suivre est de démontrer qu'il faut donner une réponse négative à cette question " (19, p. 355). It is interesting to note that Syrkin's statement is practically the opposite from Reed's—whereas Reed accepts negative correlations as indicative of convergence and rejects positive correlations as conclusive proof of divergence, Syrkin accepts the positive correlations, but casts doubt upon the negative correlations.

With an ingeniously devised example, Syrkin shows that, even in a case of " perfect divergence " in which gains are proportional to initial scores, the correlation between initial score and gain will be negative, if the reliability coefficient of the test falls below a certain value. Since Syrkin's article may not be readily available to the reader, a summary of his demonstration is reported here:

Let x and y represent the scores made by a subject in the initial and final trials, respectively, and X and Y the corresponding " true " scores, free from errors of measurement.

Let $\triangle=Y-X$ and $\delta=y-x$

Assuming perfect divergence of scores (gains proportional to initial scores), we have

$$r_{XY}=1;\ r_{\triangle X}=1;\ \text{and}\ \triangle=Y-X-qX \qquad (1)$$

in which q is a *positive factor.*

It follows that:

$$Y=(1+q)X \qquad (2)$$

$$\text{and} \quad \sigma_Y=(1+q)\sigma_X$$

The following three equations can all be readily derived from the formula for Correlation of Sums and Differences (17):

$$r_{\delta x}=\frac{r_{xy}\sigma_x\sigma_y-\sigma^2{}_x}{\sigma_x\sigma_\delta} \qquad (3)$$

$$r_{xy}\sigma_x\sigma_y=r_{XY}\sigma_X\sigma_Y \qquad (4)$$

$\sigma^2_X = \sigma^2_x P_x$ Where P_x is the reliability coefficient on the initial trial. (5)

Substituting in formula (3), values derived from formulæ (1), (4), and (5), we get

$$r_{\delta x} = \frac{P_x \sigma_X \sigma_Y - \sigma^2_X}{\sigma_x \sigma_\delta P_x} \tag{6}$$

It is obvious from formula (6) that $r_{\delta x}$ *must* be negative, even with the perfect divergence represented in this problem, whenever

$$P_x < \frac{\sigma_X}{\sigma_Y}$$

or

$$P_x < \frac{1}{1+q}$$

If, for example, the gains made are 30% of the initial score (*i.e.,* q=.30), then the correlation between initial score and gain *must* be negative, if the reliability coefficient on the initial trial is .76 or less. Likewise, the correlation will be negative if the gains are 5% of the initial score and the reliability coefficient .95 or less.

The use of tests of low reliability, without subsequent correction of the correlation coefficients for attenuation, is therefore one more factor which loads the dice in favor of the convergence interpretation.

5. *Inequality of Units.*

Data on the effects of practice upon test scores are frequently complicated by the fact that the units used to measure performance are not equal at different parts of the scale. The difference in ability required to make scores of 12 and of 13 is not necessarily the same as the difference required to make scores of 20 and 21. One unit at one part of the scale may not represent the same amount of ability as one unit at a different part. The need for equal units of measurement in mental testing is generally recognized and some efforts have been made to " scale " tests by various methods in order to secure equality of units.*

In none of the studies on practice and variability published to date, however, was any attempt made to secure an equal unit scale, although the need for such a scale was frequently recognized. Whitley, for example, described the problem as follows: " Evidently the value of such statements† would be conditioned by the nature of the test, for units near the physiological limit would not be equal to those in the lower ranges. In a test such as mental multiplication, the gain of the last few units may be far more difficult than that of the first many. In a cancellation test, the units may possibly be of rather more equal difficulty, conditioned as they are by factors of amount of eye-movement necessary, and rejection of wrong stimuli. In a feat such as juggling with balls, the first three or four units may be *harder* to gain than fifteen such units later "(**30**, p. 102). The same difficulty is suggested by Chapman, in his monograph on practice and variability (**1**, p. 17), who states that " a unit gained at the end of the practice period may be worth many units gained at the commencement." Syrkin (**19**) calls attention to the frequent neglect of this question in practice experiments and points out that in a psychological test what we are really after is equality of ability units required, not equality of either time or amount units, and that ability is not necessarily proportional to either time or amount units (**19**, pp. 353–4).

These statements are representative of what has been said regarding inequality of units in studies on practice and variability. It should be added that the use of test scores which are not expressed in terms of an equal unit scale may be especially misleading in practice experiments, more so than in other types of work. Changes in size of units have usually been found to be fairly progressive, the units at the lower and upper ends being larger than those nearer the middle of the scale. Evidence for this can be found in Thorndike's application of scaling methods to the scores

* *Cf.* for example: McCall, Wm. A., How to Measure in Education, Chs. IX and X; Thorndike, E. L., The Measurement of Intelligence, Chs. VII and IX.

† Referring to comparisons of improvement made by different subjects.

obtained on several common group intelligence tests (26, Ch. VII, pp. 224–270). Since, in the course of practice, individuals move up from one end of the scale to the other, the effect of unequal units would be of the nature of a constant error and thus yield very misleading results.

It was one of the chief purposes of the present investigation to scale the scores on the tests used, so as to express the effects of practice as nearly as possible in terms of equal units of ability. Scaling presents special difficulties when applied to practice data. All scaling techniques used so far are based upon two fundamental assumptions, viz.: (1) the greater the number of subjects who can do a thing, such as earning a particular score, the easier it is, and (2) the trait measured is normally distributed. The scaled value of each raw score is determined by the number of subjects in a large group who make that score. Obviously, in practice studies, a test cannot be scaled on the basis of the performance of the experimental group during any *one trial,* since scores on other trials would then run off scale. Any one of four methods could be used in solving this difficulty. The relative advantages and disadvantages of each method will be discussed in turn.

(1) *Scaling on the distribution obtained by throwing together the scores made by the practice subjects in all the trials:* This procedure involves two variable factors, operating simultaneously, viz., individual differences and practice effect. This in itself complicates the interpretation of the resulting curve. It is difficult to see just what the resulting curve would mean. The form of the practice curve itself would enter into the determination of the scale values. The rapid initial improvement and subsequent flattening out of the practice curve would affect the relative frequency and closeness of scores occurring at different parts of the range. The very phenomenon of practice with which the experiment deals would thus be taken into account and in part eliminated by the scaling technique.

Furthermore, the number of trials included in the distribution would have to be such as to yield a normal curve. As the limits of improvement are approached and scores begin to pile up, the distribution would no longer be normal and could not be used for scaling. Hence, a number of trials must be arbitrarily omitted from the results in order to use this scaling technique.

(2) *Scaling the scores for each trial on the basis of their own distributions, and then transmuting them all in terms of the sigma of one of the distributions:* Thurstone's absolute scaling technique, as well as that used by Thorndike in the construction of the CAVD, fall into this class. This method seems at first very well suited to practice data, but we find two obstacles in the way of its application. First, the amount of improvement from trial to trial, especially at the beginning of practice, is too great to give a sufficient amount of overlapping of adjacent distributions, required for this scaling technique. Secondly, the distributions may not be normal on all the trials. If the distributions of scores on the initial trials are normal, those in subsequent trials very frequently are not.

(3) *Scaling the test on a different and more heterogeneous group than that used in the experiment:* This is essentially the method used by McCall (8, Ch. X) in constructing the T-scale. Let us suppose, for example, that the test were scaled on a large sampling of unselected twelve-year-olds, ranging from very dull to very bright, and that the practice group consisted of eight-year-olds selected from the centre of the distribution of all eight-year-olds. The practice group would probably begin near the lower end of the twelve-year scale and might still be within the scale at the end of practice.

Practically, then, this method would probably work. Are we justified, however, in applying scale values obtained on twelve-year-olds to evaluate the improvement through practice made by eight-year-olds. Could the relative difficulty of items and the relative differences in difficulty between items vary from one age group to another? What evidence there is on this problem certainly suggests an affirmative answer. Shimberg (15) obtained different scale values for items on an information test when scaled on city and on country children. The differences in experience and general background of those two groups were sufficient to affect not only the absolute difficulty of items, but their *relative* difficulty, thus yielding different scale values. The differences in general experience and training of children in different age groups might indeed be sufficient to yield the same inconsistencies.

(4) *Scaling on a large and heterogeneous group of which the experimental group is a part:* This method is essentially a modification of the third method described above. The tests are first given to a large and heterogeneous group, and the scores scaled on the distribution obtained on this group. Then the practice group is selected from the lower end of this distribution. The scores made by the experimental subjects on *each* trial are transmuted into the scale values obtained on the large group. Thus the scores made by the practice subjects themselves on the first trial contribute towards the determination of the scale values.

The success of this method depends upon two conditions: First, the scaling group must cover such a wide range of ability that the practice subjects on their last trial will not excel the best performance of the larger group on trial 1—otherwise they would run off scale. Secondly, the original group, although heterogeneous, must not be made up of discrete groups for whom the scale values might differ. For example, the group could not be made up by throwing together two groups of widely varying age, experience, or education, and then using the poorer group as practice subjects. This procedure would be subject to the same criticism brought against the third method. Such a grouping, however, would yield either a very platykurtic or a bimodal distribution, and this would make it unsuitable for scaling purposes. The scaling group, then, although covering a wide range, should yield a normal distribution. It was our conviction that these two conditions, although offering practical difficulties, were not impossible of achievement. Accordingly, the scaling technique described in this section was chosen for use in the present investigation.

CHAPTER II

The Experimental Procedure

The general plan of the present experiment on the effects of practice upon individual differences was, first, to scale the tests chosen for the practice experiment upon a large and heterogeneous group; secondly, to use the individuals in the lower end of this group as subjects in the practice experiment, putting them through a constant number of trials in each test. The scores made by the experimental group were expressed throughout in terms of the scale constructed in step one. The details of construction and administration of tests and choice of subjects will be discussed in the following sections.

1. *Construction and Choice of Tests.*

In the choice of material to be used in the practice experiment and in the construction of the tests, the following principles were observed:

(1) Practice effect was measured by improvement in the practiced material itself, and not by special interpolated tests. This procedure has generally proved most satisfactory in measuring practice effect. Peterson and Barlow (**12**), for example, describe the disadvantages of using interpolated tests as follows: "If one uses any extraneous test of the practice effect, one is always liable to the error of using tests at the different stages of training which are not fully and equally representative of the changing functions studied, and also to the error of employing tests not equally fair in degree of difficulty at the different stages of practice."

(2) Several parallel forms of each test were constructed to be used in rotation in the successive trials, so as to avoid the memorization of specific items or sequences by the subjects. In some tests, memorization occurred less readily than in others and hence fewer parallel forms were needed in order to preclude their recognition by the subjects.

(3) The attempt was made to include items in each test which were as nearly as possible of equal difficulty throughout. Special precautions were taken, such as in the arrangement of items, to insure equality of difficulty in different forms as well as within each form.

(4) Each test form included much more material than the subject could use during the time allowed, so as not to introduce any extraneous limitations in the amount of improvement possible. This principle was followed by determining beforehand how much material adult subjects could cover when allowed a work period three or four times as long as would be allowed in the experiment proper.

(5) In order to insure full understanding of directions and familiarity with materials on the part of all subjects, a preliminary trial was given as a fore-exercise in all the tests. The procedure in this trial was identical in every way with that subsequently followed, but, of course, a different form of the test was employed. In addition to reducing individual differences in the understanding of directions, which would be an extraneous factor in the tests used, this procedure may serve to increase the reliability of the trials given subsequently. The effect of the fore-exercise upon test reliability has not been investigated very adequately as yet. Egan (3) in a study on this problem, using the National Intelligence Test with school children, reports that the use of a fore-exercise did not consistently raise the reliability coefficient of the test. The control techniques used by Egan were not, however, very adequate, since the test form *without* fore-exercise was always given two hours before the form *with* fore-exercise. The very act of taking a parallel form of the test probably served as a fore-exercise and hence the effect of the subsequent fore-exercise proper was greatly minimized. It seems reasonable to expect that reliability would be raised, however slightly, by the use of a fore-exercise. The problem may seem trivial in the testing of college students. The writer's experience in administering psychological tests to large groups of college students has shown, however, the amazing frequency with which even the simplest directions are misunderstood. The use of the

first trial as a fore-exercise was therefore considered quite essential, and accordingly only scores made on the second and subsequent trials were retained in the experiment proper.

(6) Some measure of the reliability of the tests employed is quite essential, both to insure that the tests used exceed a certain minimum of reliability and to make possible the correction of correlation coefficients for attenuation. There is much controversy regarding the relative merits of different methods of determining the reliability of a test. It has been frequently pointed out, for instance, that reliability computed by the odds and evens technique does not mean the same as that computed from retests. In practice data, however, there is no choice but to use some modification of the odds and evens technique. If scores on different trials were correlated, such a correlation would be affected by a constant practice change as well as by chance errors of measurement. In the present experiment, the procedure followed was to have the subject make a mark at a signal from the experimenter, given at the end of each quarter of the time allowed for the trial. Separate scores were then computed for each quarter and combined into two scores by adding together the two odd quarters and the two even quarters. This grouping of quarters was considered preferable to combining first and last against the middle two, since both practice and fatigue effects were operative within any one trial; the grouping used would thus yield the most comparable pair of scores from each trial.

In all, four tests were used in the present experiment. A fifth test, measuring eye-hand coördination by means of pencil mazes, was also tried, but was subsequently discarded because of scoring difficulties and low reliability. Following is a brief description of the tests finally retained, in the order in which they were administered.

(1) *Cancellation:* Five forms of this test were used in rotation. Each form consisted of 27 rows of capital letters, with 10 A's scattered in random order in each row. The position of each A was determined by dice-throwing in order to insure a purely chance arrangement. The subjects were told to make a slanting line through each A as they came to it, always working

along each row from left to right. Each subject was furnished with two blue pencils, so as to avoid special handicaps from the use of different pencils. Blue was used in preference to ordinary lead pencils since it greatly facilitated scoring. The duration of each trial was two minutes. At the end of each thirty-seconds period, subjects were told to place a cross marking off the amount of work done in the quarter.

(2) *Hidden Words:* Ten separate forms of this test were constructed, each containing seventy hidden words, seven in each row. The words were scattered in pied small type and spelled backwards. Only common four-letter English words were included. The same words were used in each form of the test, but in a different sequence. In order further to keep the difficulty of the various forms constant, the same confusion letters preceding the word proper were used with each of the words in all the forms. Subjects were told to begin at the upper right hand corner and read backwards, underlining each four-letter English word as they came to it. Blue pencils were again used. The duration of each trial was four minutes and a vertical line was placed to mark the end of each one-minute period.

(3) *Pyle Symbol-digit Substitution Test:* This was the only test not constructed especially for the present study. The standard blanks furnished by Stoelting were used. In order to minimize the memorization of order of items, the subjects were directed to begin on the left half of the blank on the odd trials and on the right half on the even trials. The duration of each trial was two minutes. The last digit written in each thirty-seconds period was circled. Each subject was furnished with No. 2 lead pencils.

(4) *Nonsense Syllable Vocabulary:* Five forms of this test were constructed, in addition to a key sheet. The key contained 50 pairs of three-letter nonsense syllables, arranged alphabetically for the first member of each pair. The test sheets contained the first syllables of all pairs, arranged in random order. The subject, with the key always before him, was to fill in the second syllable corresponding to each syllable given on the test sheet. All the syllables appeared twice on each test form. Subjects were

furnished with No. 2 lead pencils. The duration of each trial was two minutes, and subjects indicated the end of each thirty-seconds period by underlining the last syllable written.

All of the tests were scored in terms of total number of correct items completed in the given time.

2. *Administration of the Tests.*

The preliminary scaling group of 1,000 subjects was given *two* trials of each of the tests in immediate succession, in the order reported above. The tests were administered by the psychology instructors* to their respective classes during one regular class period. Standard procedure was insured by full mimeographed directions, both for the subjects and for the experimenters. The use of the first trial as a fore-exercise helped further to make conditions constant for the subsequent trials.

The practice proper was given outside of class hours to small groups of subjects at a time. All the trials were given in one sitting with one-minute rest periods between trials. Twenty trials were administered in each test except hidden words; in the latter, in view of the fact that each trial was twice as long as in the other tests, only fifteen trials were given. The subjects came by appointment on either a Thursday or a Friday afternoon at three, four, or five o'clock. Thursday and Friday of one week were devoted to each of the four tests in each institution. The practice was limited to the particular hours and days chosen in order to reduce as much as was feasible the possible effects upon performance of weekly and diurnal variations of efficiency. It would have been ideal to limit all the experimenting to one hour of one particular day, but this would have been highly impracticable in view of the number of tests, subjects, and institutions included in the present study. All of the tests given in the practice experiment were administered by the writer and two

* The writer is very grateful to all the instructors who coöperated in this project: At Columbia University Extension, Drs. P. H. DuBois and O. Klineberg; at C. C. N. Y., Prof. H. D. Marsh and Dr. G. M. Smith; at N. Y. U., Drs. E. R. Henry and R. Likert; at Brooklyn College, Drs. E. E. Asch and A. B. Wood and Mr. E. G. Brundage.

trained experimenters. The subjects were all tested in a specially chosen room in the institution which they were attending.

3. *Subjects.*

The group upon which the tests were scaled consisted of 1,000 college students of both sexes drawn from five institutions in New York City.* The subjects were all enrolled in courses in either general or experimental psychology. The composition of the group with respect to sex and institution is shown in Table III.

TABLE III

Subjects in the Scaling Group

Institution	Male	Female	Total
C. C. N. Y.	316	...	316
N. Y. U.	182	...	182
Brooklyn College	178	121	299
Barnard College	..	142	142
Columbia Extension	37	24	61
Total	713	287	1,000

The median age of the subjects was 18 years 11 months, and the average age 19–2 with an S.D. of 2½ years. The range extended from 15–6 to 33–8, although the number of cases over 25 years of age was practically negligible. As can be seen from a comparison of the mean and median values, the age distribution was skewed, with a marked piling up at the younger ages.

The subjects in the experimental group were selected from the original group of 1,000 on the basis of their scores on the preliminary trial. The selection was made independently for each test; hence the practice groups were not the same for all the tests. Many subjects, to be sure, happened to be included in the practice groups for more than one test; but this was to be expected, since the tests correlated positively with each other. The method of selection used was to call upon all subjects whose scores fell at or below —1 Q of the entire distribution. Accordingly, approx-

* The total number originally tested was 1,013, but a few papers had to be discarded because the records were incomplete or otherwise irregular. Additional papers, picked at random, were then omitted so as to bring the number down to 1,000 for each test, thereby facilitating the computation of the scale values.

imately 250 subjects were asked to come for the practice experiment in each test.* Since the subjects came for the practice trials outside of class hours, their partaking in the experiment had to be made an optional matter; hence, not all the subjects called were actually used. The number of subjects whose records were finally kept in the practice series for each test was as follows:

Cancellation	200
Hidden Words	114
Symbol-digit	134
Vocabulary	123

The subjects were paid a nominal fee for the time spent in the experiment outside of class. This monetary compensation proved a fairly good incentive. Most of the subjects also showed a keen interest in the experiment itself, and were very curious to find out their scores at the termination of the experiment and to observe their improvement through practice. No one who partook in any part of the experiment, of course, was given any indication of the method whereby the practice subjects were chosen. On the contrary, they all were led to believe that the selection was made so as to yield a random and normally distributed sampling of the original group, for more intensive experimentation. From all the evidence that could be gathered, from comments, questions, and conversation of subjects, none seemed to doubt the correctness of this statement and no disturbance was occasioned by the particular choices made.

* The number of subjects was not exactly 250 in *all* the tests, since at the beginning of the experiment a tentative value of —1 Q had to be used, based on about 200 cases. As the experiment progressed, more subjects were added to the scaling group and the value of —1 Q was determined on a progressively larger group. The differences in successive values of —1 Q obtained were, however, very small.

CHAPTER III

Evaluation of the Tests

The results bearing upon the construction of the tests will be summarized in the present chapter. All of the data reported in this chapter were obtained on the preliminary group of 1,000 subjects. In evaluating the tests used in the present investigation, the two problems which concern us most are *reliability* and *scaling*.

1. *Reliability.*

The reliability coefficients of trial two of each test on 1,000 subjects, were as follows:

Cancellation	.9098
Hidden Words	.8464
Symbol-digit	.8888
Vocabulary	.7727

These coefficients represent the reliability of the whole test and were found by applying the Spearman-Brown formula to the correlation of halves computed by the method described in Chapter II. In view of the very brief duration of the trials, the reliabilities of the tests are surprisingly high. Evidently the measures taken in the construction and administration of the tests, such as the selection and arrangement of items and the use of the first trial as a fore-exercise, were effective in yielding reliable tests. The main reason for the relatively low reliability coefficient of the vocabulary test is probably to be found in the narrower spread of scores on this test, since it did not discriminate as widely among subjects as did the other tests.

2. *Scaling the Tests.*

The first requirement for the application of the scaling technique is normality of distributions. The distributions of scores made by our group of 1,000 subjects, on trial two of each test, are given in charts I, II, III, and IV. An inspection of the curves will show how closely they approach the normal curve. Each of the four distributions was tested for normality by the method of moments. The values of β_1 and β_2, the coefficients of skewness and kurtosis, respectively, are given in Table IV.

TABLE IV

Normality of Distributions

Test	$\beta_1 \pm \text{P.E.}_{\beta_1}$	$\beta_2 \pm \text{P.E.}_{\beta_2}$
1. Cancellation	.0370±.0136	2.9528±.1130
2. Hidden Words	.1959±.0524	3.3132±.1957
3. Symbol-digit	.0516±.0199	3.0756±.1359
4. Vocabulary	.0272±.0224	3.8573±.4628

The P.E.'s of β_1 and β_2 were found from Tables 37 and 38 of Pearson's *Tables for Statisticians and Biometricians*(9). In a normal curve, β_1 is equal to zero and β_2 to 3.00 It will be seen that none of the obtained values deviates significantly (*i.e.*, by 4 or more times its P.E.) from the normal curve values. These distributions can therefore be used in the construction of the σ-scales.

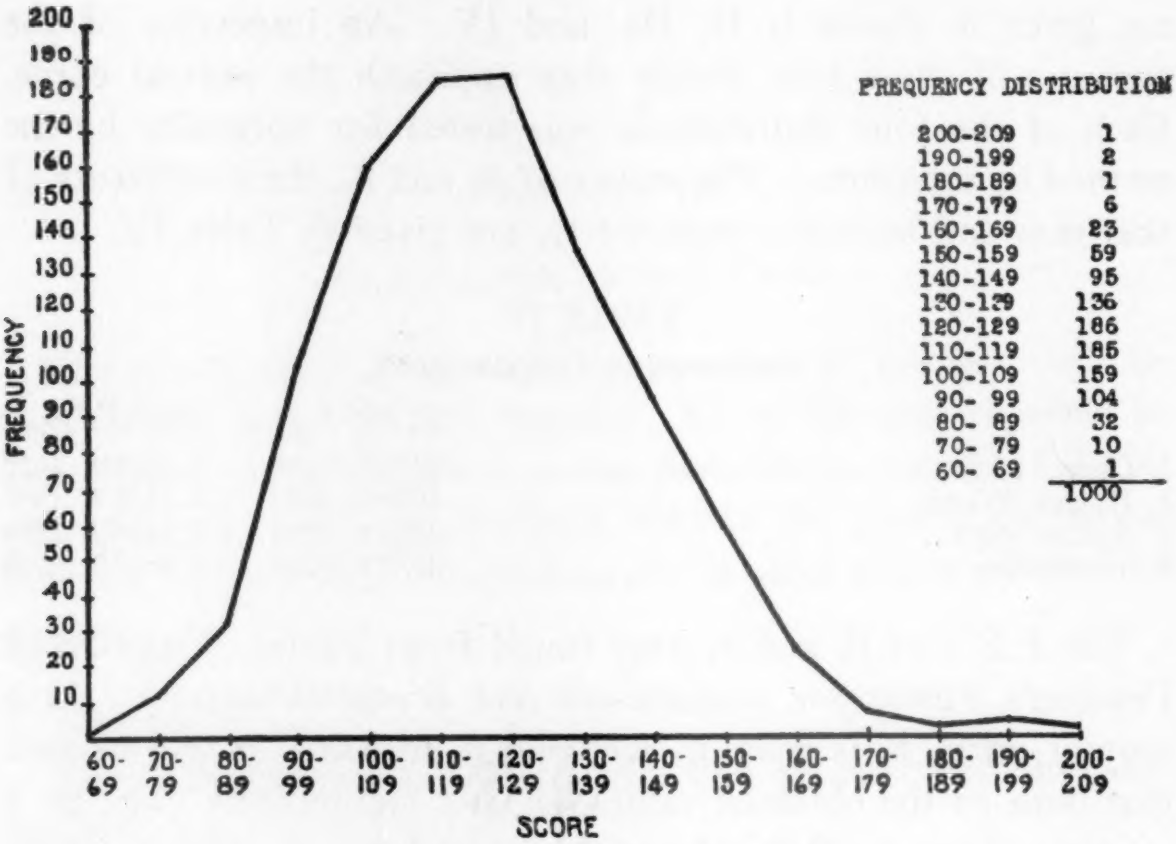
Another question which should be considered before using the data in constructing the σ-scales is the degree of variability of the scores. The success of the scaling technique employed depends upon the use of a sufficiently heterogeneous group, so that the practice subjects will not run off scale. The group tested seems to yield a satisfactory scattering of scores on each of the four tests. Following are the means, standard deviations, and range of scores in each test:

TABLE V

Central Tendency and Variability of Scaling Group

Test	Mean	S.D.	Range
1. Cancellation	120.90	20.40	65–200
2. Hidden Words	23.69	8.44	2–60
3. Symbol-digit	76.48	19.00	26–149
4. Vocabulary	32.37	5.42	8–52

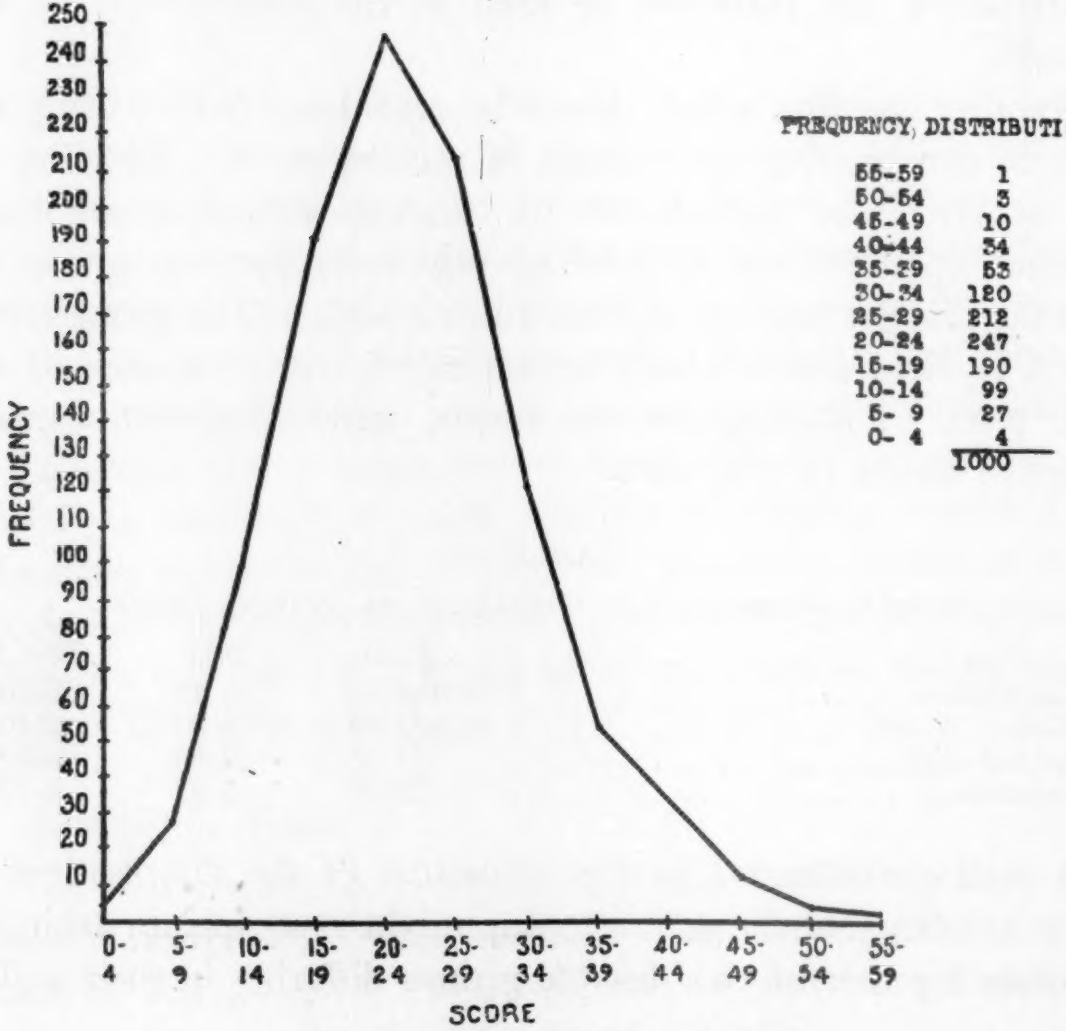
A final consideration in the evaluation of the distributions is the possibility that the scaling group might represent an arbitrary lumping together of two discrete groups differing in their ability



FREQUENCY DISTRIBUTION

200-209	1
190-199	2
180-189	1
170-179	6
160-169	23
150-159	59
140-149	95
130-139	136
120-129	186
110-119	185
100-109	159
90- 99	104
80- 89	32
70- 79	10
60- 69	1
	1000

Fig. I. Cancellation—Trial 2.



FREQUENCY, DISTRIBUTION

55-59	1
50-54	3
45-49	10
40-44	34
35-29	53
30-34	120
25-29	212
20-24	247
15-19	190
10-14	99
5- 9	27
0- 4	4
	1000

Fig. II. Hidden Words—Trial 2.

on the tests used, and that the poorer of the two groups would be used as experimental subjects. The chief factor which might yield such a discrete grouping of subjects is that of sex differences. This possibility was therefore investigated. In Table VI are presented means and standard deviations for each sex on the four tests, as well as data on the reliability of the difference between the male and female means.

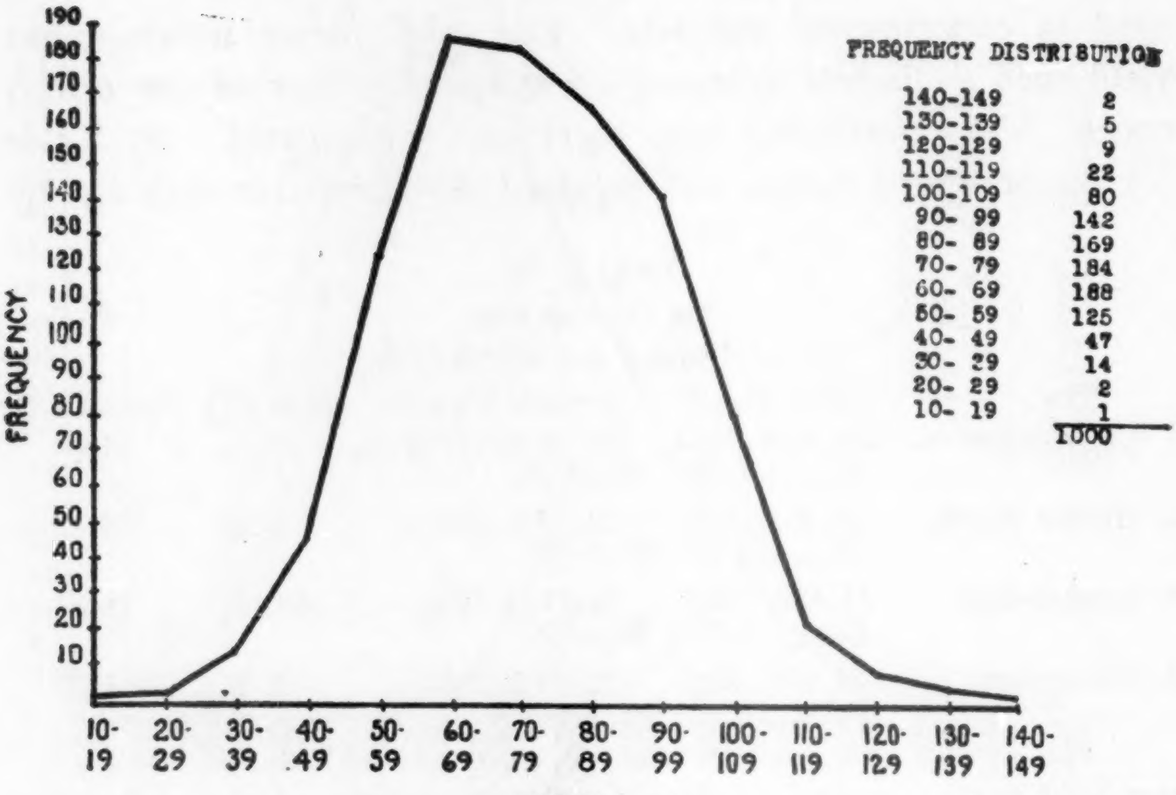
TABLE VI

SEX DIFFERENCES

(N=713 Males and 287 Females)

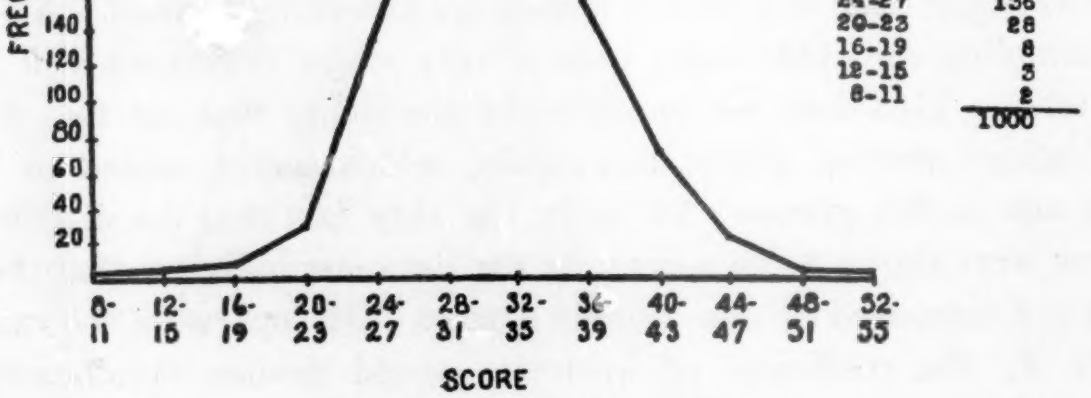
Test	Male Mean	Female Mean	Male S.D.	Female S.D.
1. Cancellation	119.43±.7702	124.59±1.1732	20.71	19.88
	$D/\sigma_D=4.56/1.40=3.26$			
2. Hidden Words	24.46±.3207	21.75±.4795	8.62	8.12
	$D/\sigma_D=2.71/.58=4.70$			
3. Symbol-digit	75.47±.7305	79.03±1.0706	19.67	18.17
	$D/\sigma_D=3.59/1.30=2.77$			
4. Vocabulary	31.97±.2057	33.34±.3080	5.50	5.20
	$D/\sigma_D=1.32/.37=3.56$			

In every case there is a reliable sex difference. Cancellation, Symbol-digit, and Vocabulary show female superiority; Hidden Words shows the males to be superior. This finding seems at first to throw doubt upon the suitability of such a sampling for the scaling technique employed. Further analysis of the data, however, brings out the fact that in spite of the reliable difference in means, the scores made by the two sexes do not show the characteristics of discrete groups. *In the first place,* the differences between the male and female means are exceedingly small. With a sampling of 1,000 cases, even a very slight difference will be reliable. This does not preclude the possibility that the two distributions overlap almost completely, which seems indeed to be the case in this group. *Secondly,* the very fact that the distributions were shown to be normal by the Beta test indicates that they are not composed of two discrete groups. If such were the case, then β_2, the coëfficient of kurtosis, would deviate significantly from 3.00. *Finally,* the sex composition of the practice group shows definitely that the method of selection employed did not yield a discrete inferior group as practice subjects. The percen-



FREQUENCY DISTRIBUTION	
140-149	2
130-139	5
120-129	9
110-119	22
100-109	80
90- 99	142
80- 89	169
70- 79	184
60- 69	188
50- 59	125
40- 49	47
30- 29	14
20- 29	2
10- 19	1
	1000

FIG. III. Symbol-Digit—Trial 2.



FREQUENCY DISTRIBUTION	
52-55	1
48-51	1
44-47	20
40-43	73
36-39	156
32-35	328
28-31	244
24-27	136
20-23	28
16-19	8
12-15	3
8-11	2
	1000

FIG. IV. Vocabulary—Trial 2.

tages of male and female subjects in the practice group for each test are given below in Table VII. The percentages of the two sexes in the original scaling group are also given for comparison.

TABLE VII

SEX COMPOSITION OF THE PRACTICE GROUPS

	Number of Cases		Per cent of Cases	
Test	Male	Female	Male	Female
1. Cancellation	161	39	80	20
2. Hidden Words	70	44	61	39
3. Symbol-digit	97	37	72	28
4. Vocabulary	96	27	78	22
Original Scaling Group (all tests)	713	287	71	29

The sex composition of the practice groups does not differ by more than 10% in either direction from that of the original scaling group. Thus the three arguments presented above show that, although the sex differences are reliable, such a finding does not invalidate the method of selection employed.

Another possibility of discrete grouping is to be found in the use of groups from various institutions. In order to investigate this possibility, Tables VIII and IX were prepared. Table VIII shows the mean score on each test found in the different institutions, as well as the mean for the entire group. Table IX shows the percentage of practice subjects drawn from each institution, in comparison with the percentage from each institution tested in the original scaling group.

TABLE VIII

MEAN SCORE OF SUBJECTS IN EACH INSTITUTION

Institution *	N	Cancellation	Hidden Words	Symbol-digit	Vocabulary
Barnard College	141	124.59	22.72	82.55	33.33
Columbia Extension	59	121.40	21.37	75.16	29.04
N. Y. U.	188	124.20	23.71	81.43	31.99
C. C. N. Y.	316	115.70	24.61	73.17	33.28
Brooklyn College	296	121.67	23.12	74.53	32.67
Total	1,000	120.90	23.69	76.48	32.37

* The institutions have been arranged in the order tested, although much of the testing went on simultaneously in more than one institution.

The means for each institution deviate as much as 5 points in either direction from the total Cancellation mean of 120.90, 2 points from the Hidden Words mean of 23.69, 6 points from the Symbol-digit mean of 76.48, and 3 points from the Vocabulary mean of 32.37. These deviations are not much greater than would result from sampling error alone. The final check of the discreteness or overlapping of the various groups lies in the comparison of the composition of the experimental group with that of the original group, shown in Table IX. It will be seen that the percentage of practice subjects chosen from each institution is quite similar to the percentage tested in the respective institutions in the original scaling group. The selection does not show any consistent deviations which might correspond to the deviations in mean scores of the various groups. The differences in percentages from test to test seem to be the results of chance factors operating to raise or lower the number of subjects coming for the experiment, and not the result of the method of selection.

TABLE IX

COMPOSITION OF PRACTICE GROUPS WITH RESPECT TO INSTITUTION

	Scaling Group		Cancellation		Hidden Words		Symbol-digit		Vocabulary *	
Institution	N	%	N	%	N	%	N	%	N	%
Barnard College....	141	*14*	16	*8*	19	*17*	16	*12*	12	*10*
Columbia Extension	59	*6*	8	*4*	4	*4*	5	*4*	10	*7*
N. Y. U...........	188	*19*	36	*18*	7	*15*	16	*12*	26	*21*
C. C. N. Y........	316	*31*	79	*39*	43	*37*	49	*36*	56	*45*
Brooklyn College...	296	*30*	61	*31*	31	*27*	48	*36*	19	*15*

* Owing to a misunderstanding, only a small number of the subjects chosen in Vocabulary at Brooklyn College were obtained. This explains the unusual percentages obtained for this test.

Thus it seemed that the distributions obtained were well suited, from every standpoint, to the application of the scaling technique chosen. Accordingly, scales were constructed upon these distributions for each of the four tests. The statistical procedure employed in constructing these scales was the same as that employed by McCall in the construction of the T-scale. The scores made by the 1,000 subjects in each test were tabulated. From these data, the percentage of subjects exceeding plus one-half of those reaching each score was found. These percentages were then used to find the σ-value of each score by reference to a table of the normal frequency surface. A copy of each of the σ-scales will be found in Tables X to XIII.

TABLE X

Cancellation Scale

Raw Score	σ-value										
181	77	162	70	143	61	124	52	105	42	86	32
180	77	161	70	142	60	123	52	104	41	85	32
179	76	160	69	141	60	122	51	103	41	84	31
178	76	159	68	140	59	121	51	102	41	83	30
177	76	158	67	139	59	120	50	101	40	82	30
176	76	157	67	138	58	119	50	100	40	81	29
175	76	156	67	137	58	118	49	99	39	80	28
174	76	155	67	136	58	117	49	98	39	79	28
173	75	154	66	135	57	116	48	97	38	78	27
172	75	153	66	134	57	115	48	96	38	77	26
171	74	152	65	133	56	114	47	95	37	76	26
170	73	151	65	132	56	113	46	94	36	75	25
169	73	150	64	131	55	112	46	93	35	74	22
168	73	149	63	130	55	111	46	92	34	73	21
167	72	148	63	129	54	110	45	91	34	72	21
166	72	147	62	128	54	109	45	90	33	71	21
165	71	146	62	127	53	108	44	89	33	70	21
164	71	145	62	126	53	107	43	88	32		
163	70	144	61	125	53	106	43	87	32		

TABLE XI

Hidden Words Scale

Raw Score	σ-value										
60	92	50	77	40	67	30	58	20	46	10	32
59	89	49	76	39	66	29	57	19	45	9	31
58	87	48	75	38	66	28	56	18	44	8	30
57	85	47	74	37	65	27	55	17	42	7	29
56	83	46	74	36	64	26	53	16	41	6	28
55	81	45	73	35	63	25	52	15	39	5	25
54	81	44	71	34	62	24	51	14	38	4	23
53	81	43	70	33	61	23	50	13	37	3	20
52	79	42	69	32	60	22	49	12	35	2	17
51	78	41	68	31	59	21	47	11	34		

TABLE XII

Symbol-Digit Scale

Raw Score	σ-value										
149	93	128	73	107	66	86	55	65	45	44	32
148	92	127	73	106	66	85	54	64	44	43	31
147	91	126	73	105	66	84	54	63	44	42	30
146	90	125	72	104	65	83	53	62	43	41	29
145	88	124	72	103	65	82	53	61	42	40	29
144	86	123	72	102	64	81	53	60	42	39	28
143	84	122	72	101	64	80	52	59	41	38	27
142	83	121	72	100	63	79	52	58	40	37	27
141	81	120	71	99	61	78	51	57	40	36	26
140	80	119	71	98	61	77	51	56	39	35	26
139	79	118	71	97	60	76	50	55	39	34	25
138	79	117	70	96	60	75	50	54	38	33	25
137	79	116	70	95	59	74	49	53	37	32	24
136	79	115	69	94	58	73	49	52	36	31	23
135	78	114	69	93	58	72	49	51	35	30	22
134	77	113	69	92	57	71	48	50	35	29	20
133	77	112	69	91	57	70	48	49	34	28	17
132	76	111	68	90	57	69	47	48	34	27	15
131	75	110	68	89	56	68	47	47	33	26	13
130	75	109	67	88	56	67	46	46	32		
129	74	108	67	87	55	66	46	45	32		

TABLE XIII

Vocabulary Scale

Raw Score	σ-value										
52	83	44	71	36	58	28	42	20	28	12	22
51	81	43	69	35	56	27	40	19	27	11	21
50	81	42	67	34	53	26	38	18	27	10	19
49	81	41	66	33	51	25	36	17	26	9	19
48	80	40	64	32	49	24	34	16	25	8	17
47	78	39	62	31	47	23	32	15	24		
46	75	38	61	30	46	22	31	14	24		
45	73	37	59	29	45	21	29	13	23		

CHAPTER IV

The Effects of Practice

All of the results reported in the present chapter were obtained on the smaller groups used in the practice series, ranging from 114 to 200 in number of subjects (cf. p. 29). The scores made by the practice subjects in each trial were first transmuted into units of the scales described in Chapter III. All of the measures reported in the present chapter were computed from scaled scores. In addition, the corresponding measures computed from raw scores are given for comparison.

1. *Central Tendency and Variability.*

The means and standard deviations for each trial are presented in Tables XIV to XVII. The values computed from raw as well as from scaled scores are given in each table. It will readily be seen from these tables that both means and standard deviations *increase* with practice in all four tests. This is true of both raw and scaled scores. The question of whether the increase in standard deviation is purely a numerical artifact, which *must* follow from the rise in mean, has already been discussed in Chapter I (cf. pp. 10–12). On the basis of the theoretical arguments presented, it was shown that this is not necessarily true. Let us now examine the data more searchingly for any experimental evidence bearing upon this question. Is the relationship between rise in mean and rise in standard deviation actually such as would be expected if the changes in standard deviation were a statistical consequence of the changes in size of mean? Are there any striking exceptions to the correspondence between changes in mean and standard deviation which would be expected from the operation of numerical factors alone? If so, can psychological factors be found whose operation might produce such a lack of correspondence? These questions may be investigated by inter-test as well as by intra-test comparisons.

Inter-test comparisons are made possible in the present study by the use of the scaling technique. The scaled scores in the various tests are all expressed in comparable units and hence direct com-

TABLE XIV

Cancellation Practice Data

(N=200)

	Raw Scores			Scaled Scores	
Trial	Mean	S.D.	Trial	Mean	S.D.
1.	101.90	11.91	1.	40.63	6.78
2.	110.20	11.91	2.	44.99	6.42
3.	113.98	12.73	3.	47.00	6.60
4.	115.58	12.24	4.	48.00	6.52
5.	121.40	13.21	5.	50.75	6.60
6.	120.55	13.14	6.	50.30	6.68
7.	123.30	13.54	7.	51.68	6.62
8.	125.55	14.42	8.	52.74	7.04
9.	126.00	14.30	9.	53.06	7.28
10.	131.93	14.93	10.	55.83	7.24
11.	129.38	14.94	11.	54.70	7.24
12.	131.15	15.48	12.	55.08	7.22
13.	132.58	16.05	13.	56.09	7.70
14.	131.58	14.61	14.	55.50	7.12
15.	136.48	15.83	15.	57.88	7.54
16.	133.88	15.95	16.	56.67	7.70
17.	134.60	15.50	17.	57.01	7.32
18.	136.08	16.09	18.	57.62	7.58
19.	134.80	15.34	19.	57.08	7.36
20.	139.98	16.52	20.	59.60	7.88

TABLE XV

Hidden Words Practice Data

(N=114)

	Raw Scores			Scaled Scores	
Trial	Mean	S.D.	Trial	Mean	S.D.
1.	18.44	5.11	1.	43.58	6.94
2.	19.18	5.33	2.	44.63	6.90
3.	22.65	5.97	3.	49.00	7.52
4.	24.42	6.41	4.	51.25	7.74
5.	27.40	6.80	5.	54.49	7.86
6.	28.07	7.28	6.	55.12	8.24
7.	31.18	8.66	7.	58.18	9.28
8.	32.89	8.29	8.	60.40	8.90
9.	33.93	8.75	9.	61.30	9.10
10.	37.07	9.91	10.	64.40	10.22
11.	35.00	10.21	11.	62.19	10.46
12.	36.04	10.50	12.	63.26	10.96
13.	39.79	11.09	13.	67.02	11.36
14.	40.86	12.18	14.	68.47	12.96
15.	42.04	10.88	15.	69.28	11.44

parisons of means and standard deviations from test to test can be made, without resort to the unwarranted use of ratio measures such as V. Table XVIII shows the changes in both means and standard deviations from the initial trial to each of the subsequent trials, in each of the four tests. The test showing the largest rise in mean from initial to final trial is Symbol-digit, whereas the largest rise in standard deviation is shown by Hidden Words. It might be objected that we are comparing the fifteenth trial of practice in Hidden Words with the twentieth in the other tests. It will be recalled, however, that the duration of each trial in Hidden Words was twice as long as in each of the other tests. If we wish to keep time spent in practice constant, we should compare the *tenth* trial in Hidden Words with the twentieth in the other tests. The results of this comparison are similar to those cited above. Hidden Words still exhibits the largest rise in standard deviation and Symbol-digit the largest rise in mean. Comparing the fifteenth trial in all the tests, we again find the same relationship.

TABLE XVI

SYMBOL-DIGIT PRACTICE DATA

(N=134)

	Raw Scores			Scaled Scores	
Trial	Mean	S.D.	Trial	Mean	S.D.
1.	59.48	12.34	1.	41.15	7.58
2.	70.52	14.02	2.	47.63	7.38
3.	80.86	14.59	3.	52.69	7.30
4.	84.93	15.88	4.	54.57	8.04
5.	91.38	16.10	5.	57.90	7.94
6.	92.99	16.63	6.	58.63	8.34
7.	98.02	17.34	7.	61.02	8.66
8.	100.11	17.28	8.	62.25	8.44
9.	103.02	16.81	9.	63.79	8.08
10.	104.25	18.79	10.	64.52	8.36
11.	106.38	16.67	11.	65.22	7.94
12.	108.84	18.72	12.	65.70	9.40
13.	110.37	17.28	13.	67.04	8.06
14.	111.34	17.95	14.	67.51	8.40
15.	111.79	17.99	15.	67.78	8.72
16.	114.40	19.37	16.	69.13	9.78
17.	113.17	17.94	17.	68.19	8.92
18.	114.59	18.29	18.	68.81	8.80
19.	115.11	17.84	19.	69.17	8.40
20.	115.67	19.91	20.	70.07	9.98

Turning to *intra-test comparisons,* we find similar discrepancies. If we take the correlation between means and standard deviations from trial to trial as a general summary of the relationship between changes in the two measures, we find that the relationship is far from perfect in most cases. The correlations for Cancellation, Symbol-digit, and Vocabulary are .79, .70, and .74, respectively; Hidden Words shows a much higher correlation of .95.

A closer examination of the changes in means and standard deviations brings out several major discrepancies. The most striking discrepancy is found in the initial stages of practice. In each test, the means show a rapid rise from trials one to two; in addition, the rises during the first few trials tend to be greater than those occurring later. These findings agree with those of most investigators. The standard deviations, on the other hand, show in every test a slight initial *drop* from trials one to two. Following this initial drop, slight increases in standard deviation are found for a few trials, the largest rises usually appearing only when the later trials are reached.

TABLE XVII

Vocabulary Practice Data

(N=123)

	Raw Scores			Scaled Scores	
Trial	Mean	S.D.	Trial	Mean	S.D.
1.	26.63	3.89	1.	39.06	6.84
2.	30.61	3.29	2.	46.30	6.03
3.	30.02	3.55	3.	45.22	6.95
4.	31.48	3.18	4.	47.74	5.88
5.	31.98	3.64	5.	49.19	6.86
6.	31.93	3.70	6.	48.80	6.78
7.	33.63	3.83	7.	52.06	7.22
8.	32.00	4.30	8.	48.97	7.89
9.	33.30	3.80	9.	51.59	7.16
10.	33.87	4.41	10.	52.50	8.34
11.	34.15	4.77	11.	53.08	8.90
12.	35.39	4.40	12.	55.35	8.10
13.	34.29	4.23	13.	54.54	7.98
14.	35.11	3.85	14.	54.74	7.26
15.	35.70	4.62	15.	56.02	8.49
16.	36.03	4.59	16.	56.48	8.46
17.	36.87	4.73	17.	57.83	8.59
18.	36.17	4.92	18.	56.63	9.13
19.	36.57	4.79	19.	56.97	8.89
20.	37.57	4.88	20.	59.28	8.87

In Cancellation, the standard deviation in trials two to seven, inclusive, is *lower* than that in trial one, the steady and marked increase setting in as late as the eighth trial. In Hidden Words, the standard deviation drops slightly on trial two; from the third trial on, the standard deviation shows a fairly steady rise with only two major exceptions, viz., in trials eight and fifteen, in which the standard deviation drops although the mean continues to rise. In Symbol-digit, the standard deviation decreases on the second and third trials, although by far the largest single rises in mean take place in these two trials. In Vocabulary, the standard deviation rises and falls irregularly until the eighth trial, when the first large rise occurs, although in this trial the mean actually drops.

TABLE XVIII

Gain from Initial Trial to Each Subsequent Trial

	Cancellation		Hidden Words		Symbol-digit		Vocabulary	
Trial 1 to:	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
2.	4.36	—.36	1.05	—.04	6.48	—.20	7.24	—.81
3.	6.37	—.18	5.42	.58	11.54	—.28	6.16	.11
4.	7.37	—.26	7.67	.80	13.42	.46	8.68	—.96
5.	10.12	—.18	10.91	.92	16.75	.36	10.13	.02
6.	9.67	—.10	11.54	1.30	17.48	.76	9.74	—.06
7.	11.05	—.16	14.60	2.34	19.87	1.08	13.00	.38
8.	12.11	.26	16.82	1.96	21.10	.86	9.91	1.05
9.	12.43	.50	17.72	2.16	22.64	.50	12.53	.32
10.	15.20	.46	20.82	3.28	23.37	.78	13.44	1.50
11.	14.07	.46	18.61	3.52	24.07	.36	14.02	2.06
12.	14.45	.44	19.68	4.02	24.55	1.82	16.29	1.26
13.	15.46	.92	23.44	4.42	25.89	.48	15.48	1.14
14.	14.87	.34	24.89	6.02	26.36	.82	15.68	.42
15.	17.25	.76	25.70	4.50	26.63	1.14	16.96	1.65
16.	16.04	.92			27.98	2.20	17.42	1.62
17.	16.38	.54			27.04	1.34	18.77	1.75
18.	16.99	.80			27.66	1.22	17.57	2.29
19.	16.45	.58			28.02	.82	17.91	2.05
20.	18.97	1.10			28.92	2.40	20.22	2.03

Another discrepancy between changes in mean and standard deviation is to be found in the presence of end-spurts. All of the tests show an end-spurt in mean on the last trial, as is usually found in work curves. The standard deviations, however, do not show such an end-spurt, and in fact, in Hidden Words and Vocabulary the standard deviation on the final trial is *lower* than that on preceding trials.

We thus get a picture of the mean performance on each test showing a rapid initial rise and gradual diminution of improvement, with an end-spurt on the final trial which suggests that the subjects, aware that it was the last trial, put forth extra effort to better their preceding records. Variability, on the other hand, presents a somewhat different picture. On the initial trial of the practice period, the subjects were becoming adjusted to the novelty of the situation. Hence individual differences in susceptibility to such adjustment, in response to a relatively unfamiliar situation, or quickness in " settling down " to the task, entered in as an additional factor making for variability. This might account for the slightly greater variability found in all the tests on the first trial as contrasted with the trials immediately following. The delay in the appearance of rapid rises in variability might indicate that subjects improve more uniformly during the relatively easy stages of learning, and that only when improvement becomes more difficult do we find individual differences in rate of improvement entering in to increase variability.

The absence of an " end-spurt " rise in variability to correspond to that in the mean may suggest that the subjects were affected in a fairly uniform manner by the " end-spurt " effect. This is in accord with our expectations. All of the subjects seemed highly motivated throughout the experiment, trying almost frantically to better their performance during each trial. Since very few errors were made, the subjects could estimate roughly how well they had done during each trial from the amount of material covered. On the final trial, realizing that it was the last chance to raise their scores, the subjects concentrated all their efforts toward that end.

The effect of the " end-spurt " can be seen more clearly in those tests in which the subjects seemed to be reaching a practice limit during the last few trials. In Cancellation and Vocabulary, the means of the scaled scores fluctuated irregularly between 56 and 57 during trials fifteen to nineteen and then showed an unmistakable rise on trial twenty. An examination of the individual records in these two tests brings out the relative uniformity of the operation of the " end-spurt " effect. Of the 200 subjects who practiced Cancellation, 78 show a rise in score from trials eighteen to nineteen, whereas 140 show the " end-spurt " rise on trial twenty. Similarly, 58 of the 123 subjects who practiced Vocabulary show improvement from trials eighteen to nineteen and 74 improve on the last trial. The " end-spurt ", then, was not characteristic of only a few subjects who might have unduly raised the mean, but was quite general, manifesting itself in much more than half of the cases. Since the " end-spurt " affected more subjects than did the ordinary improvement through practice in preceding trials, we should expect it to raise the mean without increasing variability, and this is, indeed, exactly what we do find.

The above analysis suggests that the observed changes in standard deviation are chiefly the result of the operation of psychological factors, rather than being mere numerical artifacts. Our analysis, to be sure, is not to be regarded as anything more than a plausible hypothesis which happens to fit the observed data. But the evidence is suggestive, even if far from conclusive, in pointing out that conditions other than mere size of average may account for changes in standard deviation. Size of standard deviation is not perfectly correlated with size of mean. Whenever the operation of psychological factors would be expected to affect means and standard deviations differently, it is actually found that one may rise and the other drop simultaneously. Thus it seems plausible to conclude that the rise in standard deviation from early to later trials found in the present experiment may be indicative of a *true* increase in individual differences with practice.

2. *Correlations.*

Product-moment coefficients of correlation were computed between (1) initial and final score in each test and (2) initial score and gain. These correlations were computed for both scaled and raw scores. The correlation coefficients obtained with scaled scores, which are the ones to be used in the final evaluation of results, were corrected for attenuation. The correlations between initial and final score were corrected by the commonly used formula derived by Spearman (16):

$$r_{xy} = \frac{r_{x_1y_1}}{\sqrt{r_{x_1x_2}}\ \sqrt{r_{y_1y_2}}}$$

In order to correct the correlations between initial score and gain for attenuation, Thomson's formula (21) was used:

$$r_{ag} = \frac{\sigma_z r_{xz} - \sigma_x r_x}{\sqrt{r_x(r_x\sigma^2_x + r_z\sigma^2_z - 2\sigma_x\sigma_z r_{xz})}}$$

in which r_{ag} is the true r between initial score and gain, x and z are the initial and final scores, and r_x and r_z their respective reliability coefficients.

1. *Reliability Coefficients.* The reliability coefficients of initial and final trials in each test were computed by correlating the sum of the scores on the odd quarters with that of the scores on the even quarters and then applying the Spearman-Brown formula. These coefficients, as well as the number of cases upon which each was computed and the standard deviations of the respective distributions, are presented in Table XIX below.

It will be noted that the reliability coefficients of the tests tend to increase with practice. It also appears, in general, that the increase in reliability is greater in those tests in which the effects of practice upon mean and standard deviation are greater. Hidden Words, with the second largest rise in mean and the largest rise in standard deviation, shows by far the largest rise in reliability with practice. Cancellation shows a smaller rise and Vocabulary a negligible drop of .03. In both of these tests the

TABLE XIX

RELIABILITY COEFFICIENTS OF INITIAL AND FINAL TRIALS

		Initial Trial		Final Trial	
Test	N	Reliability	S.D.	Reliability	S.D.
1. Cancellation	200	.7662	6.78	.8981	7.88
2. Hidden Words	114	.5423	6.94	.8461	11.44
3. Symbol-digit	134	.8422	7.58	.8498	9.98
4. Vocabulary	123	.7812	6.84	.7483	8.87

practice effect was less marked than in Hidden Words. The negligible rise in the reliability coefficient of Symbol-digit may be owing to the fact that this test began with the relatively high reliability of .8422 on the first trial, thus making further improvement in reliability very difficult.

The reliability coefficients of each test, computed on the practice subjects, may be compared with those obtained on the larger group of 1,000 subjects and reported in Chapter III. Since the size of the standard deviation in both groups is known, the reliability coefficient to be expected on the smaller group can be estimated from that of the larger group by the following formula (5):

$$\frac{\sigma}{\Sigma} = \frac{\sqrt{1-R_{11}}}{\sqrt{1-r_{11}}}$$

The estimated values of the reliability coefficients are presented in Table XX. In each test, Σ, or the standard deviation of the group of 1,000 subjects, is 10, since in the scaling process all scores were transmuted into distributions with a mean of 50 and a standard deviation of 10.

The discrepancies between the estimated and the obtained values of the reliability coefficients suggest that the tests are not equally reliable throughout the range. Cancellation and Hidden Words are a little less reliable in the lower range of scores covered by our practice subjects than they are in the upper parts of the range. Symbol-digit and Vocabulary, on the other hand, seem to be more reliable for the poorer subjects. This is strikingly true of Vocabulary. Both of these tests were such as to involve speed of writing to a much greater extent in the case of

TABLE XX

Estimated and Obtained Reliability Coefficients on Practice Groups

Test	Reliability on Group of 1,000	Estimated Reliability on Practice Group	Obtained Reliability on Practice Group
1. Cancellation........	.9098	.8042	.7662
2. Hidden Words	.8464	.6801	.5423
3. Symbol-digit.......	.8888	.8064	.8422
4. Vocabulary........	.7727	.5141	.7812

the better subjects than in the case of those making lower scores. This may have introduced a spurious factor which varied irregularly in different quarters of the trial and hence served to reduce reliability for the better subjects. With the poorer subjects, the substitution reaction itself required more time and hence variability in speed of writing did not enter in as often. Corroboration of this interpretation is to be found in the fact that the same tests which yield a higher reliability in the lower part of the range, viz., Symbol-digit and Vocabulary, show in one case no change and in the other a slight drop in reliability as the scores rise with practice.

TABLE XXI

Uncorrected Correlation Coefficients Between Initial and Final Scores

Test	r of Scaled Scores	r of Raw Scores
1. Cancellation	.5578	.5613
2. Hidden Words	.5580	.6078
3. Symbol-digit	.2525	.2972
4. Vocabulary	.3879	.3925

2. *Correlation between Initial and Final Score.* Table XXI gives the uncorrected correlations between initial and final scores computed from scaled and from raw scores. In Table XXII are shown the correlations of the scaled scores corrected for attenuation. The P.E.'s of each of the correlation coefficients reported in Table XXII have been computed by the following formula(6), which gives the P.E. of an r that has been corrected for attenuation:

$$\text{P.E.}_{r_{\infty\infty}} = \frac{.6745\ r_{\infty\infty}}{\sqrt{N}}\left\{r^2_{\infty\infty} + \frac{1}{r^2_{12}} + \left(\frac{1}{4r^2_{1I}} - \frac{r^2_{1I}}{4} + r_{1I} - 1\right)\right.$$
$$\left. + \left(\frac{1}{4r^2_{2II}} - \frac{r^2_{2II}}{4} + r_{2II} - 1\right)\right\}^{\frac{1}{2}}$$

All of the correlations presented in Table XXII are positive and significant when evaluated in terms of their P.E.'s. Most of the correlations are very high. This finding shows a marked tendency for subjects to maintain the same relative position in the group from the beginning to the end of the practice series. This does not, to be sure, have any bearing upon the question of increase or decrease in variability with practice, since it is quite possible for the initially better subjects to improve much less than the initially poorer and still maintain their initial *position* in the group. Even a perfect positive correlation between initial and final standing is compatible with a marked decrease in variability with practice.

TABLE XXII

CORRELATION COEFFICIENTS BETWEEN INITIAL AND FINAL SCALED SCORES, CORRECTED FOR ATTENUATION

Test	Corrected r	P.E.r
1. Cancellation	.6725	.0363
2. Hidden Words	.8239	.0565
3. Symbol-digit	.2981	.0645
4. Vocabulary	.5073	.0642

3. *Correlation between Initial Score and Gain.* Table XXIII shows the raw correlations between initial scores and gains from initial to final trial on each test. The correlations have been computed for both scaled and raw scores. All of these correlations are negative and fairly high. This is a very common finding and has been reported almost unanimously by previous investigators. A different picture is presented, however, by the correlations which have been corrected for attenuation. These correlations are given in Table XXIV.

TABLE XXIII

RAW CORRELATIONS BETWEEN INITIAL SCORES AND GAINS

Test	r of Scaled Scores	r of Raw Scores
1. Cancellation	—.3477	—.1874
2. Hidden Words	—.0239	—.1643
3. Symbol-digit	—.4638	—.3479
4. Vocabulary	—.3878	—.4026

In Hidden Words, in which speed of writing played a very insignificant rôle throughout the practice series, the corrected correlation between initial scores and gains is positive and high. The initially better subjects in this test tended to make the largest gains from the first to the last trial. In the remaining three tests, the correlations are still negative, but considerably lower than were the corresponding raw correlations. In these three tests, it will be remembered, speed of writing played a much greater part than in Hidden Words. Such a condition would certainly place a restriction upon the improvement of the better subjects beyond a certain point. It is interesting to find that the negative correlation is much lower in Cancellation than in either Symbol-digit or Vocabulary. Success in the Cancellation test, from the very beginning, depends largely upon speed of writing movements. Hence the better subjects in this test are just those who are initially better in speed of writing and who would be expected to show the most improvement in this activity with practice. In Symbol-digit and Vocabulary, on the other hand, the initially best subjects in the practice group are not necessarily those with the greatest speed of writing, since the ability to write rapidly does not play a large part in performance on these tests during the early stages of practice. The process of substitution itself takes so long at this stage that it obscures individual differences in speed of writing. In the later stages of practice, however, speed of writing comes to play a progressively larger part in determining individual differences in score, and hence the negative correlation between initial scores and gains.

TABLE XXIV

Correlations Between Initial Scores and Gains Corrected for Attenuation

Test	Corrected r
1. Cancellation	—.1631
2. Hidden Words	+.5122
3. Symbol-digit	—.4316
4. Vocabulary	—.3096

The four tests used in this experiment bring out quite clearly the varying results that may be expected when initial scores and gains are correlated in different types of tests. Hidden Words, which measures a relatively more complex process than any of the other tests and is a more difficult test for the subjects, yields a high positive correlation between initial scores and gains. Cancellation, measuring a more highly motor and simpler task, in which improvement is limited by physical factors, gives a low negative correlation. The remaining two tests, Symbol-digit and Vocabulary, undergo a change in the course of repetition which is probably very common in experiments on practice. Starting out as tests in which individual differences depend chiefly upon such factors as speed of forming associations and other psychological processes, they end up largely as tests of motor speed. This is just the type of situation which would produce a negative correlation between initial scores and gains.

3. *The Effect of Scaling.*

Most of the results reported in the present chapter have been given in terms of raw as well as scaled scores. It may be well to compare the two sets of results. In a very general way, the same sort of conclusions could have been reached with raw as with scaled data. In both cases, the standard deviations increase with practice. The uncorrected correlations are also roughly similar for both sets of scores; the correlations between initial and final scores are positive in both cases; those between initial scores and gains are negative in both. This is, of course, to be expected, and it is the usual result when scaled and raw data are compared. For many purposes, when only general trends are sought, the refinements of scaling may be superfluous. If a more detailed analysis of results is desired, however, the differences which result from scaling cannot be ignored. The part which scaling has played in the present study may be summarized as follows:

1. Through the use of the scaling technique, the scores on each of the four tests were expressed in the same units, since they had all been transmuted into distributions with a mean of 50 and a standard deviation of 10. This made it possible to make inter-test comparisons which would have otherwise been impossible without the use of ratios.

2. Comparisons of changes in mean and standard deviation from trial to trial, other than the gross general trend towards rise with practice in both measures, could not have been made with raw scores. The other relatively slight changes would be obscured by the use of a crude measuring instrument with unequal

units. The initial drop in standard deviation on trial two, for example, is not consistently found in the raw data. The delay in the point of maximum rise as well as the absence of an " end-spurt " in standard deviation would not have been brought out so distinctly in the raw data.

3. Finally, it should be recalled that the tests were originally constructed with the aim of getting equal units. Every care was exercised to achieve this end. The very fact that the scaled and the raw data do not differ any more than they do testifies to the conclusion that this aim was to a large extent achieved. Without the application of the scaling technique, however, equality of units could not have been assumed *a priori*. For many tests in common use, the assumption obviously does not hold. In the present investigation, since the scores *were* actually scaled, it is certain that inequality of units at different parts of the range could not spuriously have produced the results obtained.

CHAPTER V

SUMMARY

1. A survey of the literature on practice and variability revealed marked controversy on certain fundamental methodological issues. The analysis of these issues, with special reference to the mathematical assumptions involved, has pointed to the conclusion that it is theoretically sounder to use amount scores rather than time scores, and to express variability in terms of absolute rather than relative measures. Either the time constant or the amount constant method may be used, provided the experimenter states clearly at the outset what he means by equal amounts of practice. The time constant method seems, however, preferable in actual practice, since it facilitates the expression of scores in terms of amount of work per unit of time. Correlations between initial score and gain may be used if interpreted with caution. Especially important in this connection is the consideration of errors of measurement, which do not affect these correlations in the same manner as they do other correlations, but may change a positive correlation into a negative one or raise the numerical value of a low negative correlation. The available formulæ for correcting such correlations for attenuation should be employed before final conclusions are drawn. Finally, the use of a scale of equal units, the need for which has been generally recognized by mental testers, is of especial importance in studying the effects of practice upon individual differences. Of several possible scaling techniques, the one which seems preferable for use on practice data is the scaling of test scores on a large and heterogeneous group of which the experimental group is an integral part.

2. Four tests, Cancellation, Hidden Words, Symbol-digit, and Vocabulary, were scaled on 1,000 college students of both sexes. The subjects in this group whose scores fell at or below —1 Q

of the total distribution were selected for the practice experiment. The practice consisted of 15 trials in Hidden Words, and 20 trials in each of the other tests. All of the trials of one test were administered at a single sitting.

3. Individual differences, measured by the standard deviation, increased from the first to the last trial in each test. The use of the standard deviation instead of some other measure as an index of changes in variability with practice is theoretically justifiable. In addition, evidence was presented which suggested strongly that the rise in standard deviation with practice is not a mere statistical artifact which results necessarily from the rise in size of scores with practice. The correlations computed show that individuals tend to maintain the same relative positions during practice; and, unless prevented by extraneous limitations of improvement, the better subjects tend to improve somewhat more than the poorer subjects.

Bibliography

1. Chapman, J. C. Individual differences in ability and improvement and their correlations. *Teachers College, Columbia University, Contributions to Education,* 1914, No. 63, pp. 45.
2. Chapman, J. C. Statistical considerations in interpreting the effect of training on individual differences. *Psychological Review,* 1925, **32**, 224–234.
3. Egan, Eula Pearl. The effect of fore-exercises on test reliability. *George Peabody College for Teachers,* 1932, pp. 37.
4. Hollingworth, H. L. Individual differences before, during and after practice. *Psychological Review,* 1914, **21**, 1–8.
5. Kelley, T. L. The reliability of test scores. *Journal of Educational Research,* 1921, **3**, 370–379.
6. Kelley, T. L. Statistical method. N. Y.: Macmillan, 1924.
7. Kincaid, Margaret. A study of individual differences in learning. *Psychological Review,* 1925, **32**, 34–53.
8. McCall, W. A. How to measure in education. N. Y.: Macmillan, 1923.
9. Pearson, Karl. Tables for statisticians and biometricians. Cambridge University Press, 1914.
10. Peterson, Joseph. Experiments in ball-tossing: The significance of learning curves. *Journal of Experimental Psychology,* 1917, **2**, 178–224.
11. Peterson, Joseph. Thurstone's measures of variability in learning. *Psychological Bulletin,* 1918, **15**, 452–456.
12. Peterson, Joseph, and Barlow, M. C. The effects of practice on individual differences. *The Twenty-Seventh Yearbook of the National Society for the Study of Education,* Part II, 1928, 211–230.
13. Reed, H. B. The effect of training on individual differences. *Journal of Experimental Psychology,* 1924, **7**, 186–201.

14. Reed, H. B. The influence of training on changes in variability in achievement. *Psychological Monographs,* 1931, **41**, pp. 59.

15. Shimberg, M. E. An investigation into the validity of norms, with special reference to urban and rural groups. *Archives of Psychology,* 1929, **104**, pp. 84.

16. Spearman, Carl. The proof and measurement of association between two things. *American Journal of Psychology,* 1904, **15**, 72–101.

17. Spearman, Carl. Correlations of sums or differences. *British Journal of Psychology,* 1913, **5**, 417–426.

18. Stoddard, G. D. The problem of individual differences in learning. *Psychological Review,* 1925, **32**, 479–485.

19. Syrkin, M. La question de la " convergence " ou de la " divergence " sous l'aspect de la variabilité fluctuante. *Revue de la Science du Travail,* 1930, **2**, 353–364.

20. Thomson, G. H. A formula to correct for the effect of errors of measurement on the correlations of initial values with gains. *Journal of Experimental Psychology,* 1924, **7**, 321–324.

21. Thomson, G. H. An alternative formula for the true correlation of initial values with gains. *Journal of Experimental Psychology,* 1925, **8**, 323–324.

22. Thorndike, E. L. The effect of practice in the case of a purely intellectual function. *American Journal of Psychology,* 1908, **19**, 374–384.

23. Thorndike, E. L. An introduction to the theory of mental and social measurements. Teachers College, Columbia University, 1913.

24. Thorndike, E. L. Educational psychology. Teachers College, Columbia University, 1914, Vol. III, Part II.

25. Thorndike, E. L. The influence of the chance imperfections of measures upon the relation of initial score to gain or loss. *Journal of Experimental Psychology,* 1924, **7**, 225–232.

26. Thorndike, E. L. The measurement of intelligence. Teachers College, Columbia University, 1926.

27. Thurstone, L. L. A method of scaling psychological and educational tests. *Journal of Educational Psychology,* 1925, **16**, 433–451.

28. Thurstone, L. L. The unit of measurement in educational scales. *Journal of Educational Psychology,* 1927, **18**, 505–524.

29. Wells, F. L. The relation of practice to individual differences. *American Journal of Psychology,* 1912, **23**, 75–88.

30. Whitley, M. T. An empirical study of certain tests for individual differences. *Archives of Psychology,* 1911, **19**, pp. 146.

PSYCHOLOGICAL MO[illegible]RAPHS

VOL. XXXIV

[illegible] 15[illegible] [illegible]uence of Tuition in the Acquisition a[illegible] [illegible]ED LIEN WANG. Pp. 61. [illegible] in [illegible] Splitting the Mind: An Experimental [illegible] of Normal Men. CHARLES T. [illegible] Pp. 132. $2.00. 156. Relation of the [illegible] Response to Intelligence. J. A[illegible]RTH. Pp. 62. $.50. 157. An Experimental [illegible] Retention and its Relation [illegible]gence. AUG LANIER LEE. Pp. 45. $.75. [illegible] Energy Value of the Mini[illegible]ible Chromatic and Achromatic for Differe[illegible] Lengths of the Spectrum. MA[illegible] MONROE. Pp. 60. $1.00. 159. An Expe[illegible] Study of Some Behavior $1.[illegible] the Influence of the Factor of Intelligence [illegible] of the Learning Curve. GU[illegible] BUCH. Pp. 64. $1.00. 161. Further [illegible] in Retroactive Inhibition. ED[illegible]ON SKAGGS. Pp. 60. $.50.

VOL. XXXV

[illegible] Melodic Relativity of Tones. OTTO OR[illegible] [illegible] 163. Studies in Psych[illegible] the University of Illinois. Pp. 151. [illegible] Comprehensive Units in Le[illegible]pewriting. J. W. BARTON. Pp. 45. $.75. [illegible] Experimental Study of the [illegible] Psychology. ELISABETH WHEELER AMEN. [illegible] $.75. 166. A Study of the [illegible]formance of American, Mexican, and N[illegible]ren. HELEN LOIS KOCH and [illegible] SIMMONS. Pp. 116. $1.75.

VOL. XXXVI

16[illegible] [illegible]sity of Iowa Studies in Psychology. No. [illegible] by CARL E. SEASHORE. Pp. [illegible].00. 168. University of Iowa Studies in [illegible]y, No. X. Edited by CA[illegible]SHORE. Pp. 114. $1.75. 169. Attitude in [illegible] to Learning. HELEN E. SUL[illegible]p. 189. $2.75.

VOL. XXXVII

17[illegible] [illegible]reative Imagination of Théophile Gautier[illegible] [illegible]y in Literary Psychology. LOU[illegible]ILLINGHAM. Pp. 355. $4.50. 171. An E[illegible] [illegible]tal Study of Effects and Asso[illegible] [illegible]e to Certain Odors. J. H. KENNETH. [illegible] 64. $1.00. 172. Individual [illegible] Imagery. CHARLES H. GRIFFITTS. Pp. [illegible]

VOL. XXXVIII

17[illegible] [illegible]y of the Purkinje Phenomenon. LOUISE [illegible] Pp. 86. $1.25. 174. A Qua[illegible] Study of Chromatic Adaptation. MARY [illegible] ALMACK. Pp. 115. $1.75. 175. [illegible]ental Analysis of the Sensori-Motor Co[illegible] [illegible]s of Passive Oscillation: Ro[illegible] Rectilinear. ROLAND C. TRAVIS and RA[illegible] [illegible]DGE. Pp. 96. $1.40. 176. Uni[illegible] Iowa Studies in Psychology. No. XI. [illegible] CHRISTIAN A. RUCKMICK. Pp. [illegible]0.

VOL. XXXIX

17[illegible] [illegible]al Destruction in Its Relation to Maze Le[illegible] NORMAN CAMERON. Pp. 65. $1.0[illegible] [illegible]niversity of Iowa Studies in Psychology[illegible] (Seashore Commemorative Vol[illegible] [illegible]dited by WALTER R. MILES and DANIEL [illegible] Pp. 223. $3.50. 179. Time Pat[illegible] Prose. HELEN GRIFFITH. Pp. 82. $1.[illegible] An Experimental Study of Re[illegible] and Recall in Abnormal Mental Cases. [illegible] WYLIE. Pp. 81. $1.25.

VOL. XL

1[illegible] [illegible]sity of Iowa Studies in Psychology. N[illegible] Edited by CHRISTIAN A. RU[illegible] Pp. 214. $3.50. 182. Common Annoy[illegible] [illegible] Psychological Study of Ev[illegible] [illegible]versions and Irritations. HULSEY CASO[illegible] [illegible]18. $3.25. 183. A Quantitat[illegible] of Achromatic and Chromatic Sensitivit[illegible] Center to Periphery of the Vi[illegible] HAZEL A. WENTWORTH. Pp. 189. $2.7[illegible]

VOL. XLI

[illegible] [illegible]ioned Eyelid Reactions to a Light Stim[illegible] [illegible]ed on the Reflex Wink. [illegible] So[illegible] [illegible]EST R. HILGARD. Pp. 50. $.75. 185. Th[illegible] [illegible]e of Training on Changes in [illegible]y in Achievement. HOMER B. REED. [illegible] [illegible]5. 186. Judging Emotions fro[illegible] Expressions. LEO KANNER. Pp. 91. $[illegible] University of Iowa Studies in [illegible]y. No. XIV. Edited by CHRISTIAN A. [illegible] Pp. 327. $4.00.

VOL. XLII

[illegible] [illegible]cation of the Lid-Reflex by Voluntarily [illegible]ets. HELEN PEAK. Pp. 6[illegible] $1[illegible] [illegible]n Investigation of the Responses Made [illegible] [illegible]g a Multiple Choice Ma[illegible] Fr[illegible] H. LUMLEY. Pp. 61. $1.00. 190. An Exp[illegible] Study of Apparent Move[illegible] m[illegible] MIBAI. Pp. 91. $1.50. 191. An Object[illegible] [illegible]y of Concept Formation. K[illegible] SMOKE. Pp. 46. $.75. 192. General [illegible] [illegible]fic Attitudes. [illegible] CA[illegible]p. 199. $1.[illegible]. 193. An Experimental St[illegible] [illegible]ch Recognition. LAURENCE A. [illegible] Pp. 124. $1.75.

VOL. XLIII

1[illegible] [illegible]rsity of Iowa Studies in Psychology. [illegible] Edited by CHRISTIAN A. RU[illegible] Pp. 308. $3.75. 195. A Study of a Com[illegible] [illegible]nd Its Groups and Institutions [illegible]ved of as Behaviors of Individuals. RI[illegible] SCHANCK. Pp. 133. $2.00. 19[illegible] [illegible]perimental Investigation of Eye Movem[illegible] Learning to Spell Words. L[illegible] GILBERT. Pp. 81. $1.50.

VOL. XLIV

[illegible] [illegible]rsity of Kansas Studies in Psycholog[illegible] [illegible] Edited by RAYMOND H. W[illegible] Pp. 300. $3.75. 198. University of Iow[illegible] [illegible] in Psychology. No. XVI. [illegible] CHRISTIAN A. RUCKMICK. Pp. 88. $1.50. [illegible]versity of Iowa Studies in [illegible] No. XVII. Edited by CHRISTIAN A. RU[illegible] Pp. 121. $1.75.

VOL. XLV

[illegible] [illegible]rsity of Iowa Studies in Psychology. N[illegible] (Studies in the Psychology o[illegible] [illegible]ted by NORMAN C. MEIER). Edited by [illegible]AN A. RUCKMICK. Pp. 1[illegible] [illegible] An Experimental Study of Factors [illegible]g Consonance [illegible] [illegible] BUGG. Pp. 100. $1.50. 202. Binocular [illegible]ocular Relations in [illegible] [illegible]lation. T. W. COOK. Pp. 86. $1.25. 2[illegible] Fields and the [illegible] [illegible]. A. GENGERELLI. Pp. 115. $1.50.

Directory of American Psychological Periodicals

American Journal of Psychology—Ithaca, N. Y.: Cornell University.
Subscription $6.50. 624 pages ann. Ed. by M. F. Washburn, Madison Bentley, K. M. Dallenbach and E. G. Boring.
Quarterly. General and experimental psychology. Founded 1887.

Journal of Genetic Psychology—Worcester, Mass.: Clark University Press.
Subscription $14.00 per year; $7.00 per vol. 1000 pages ann. (2 vols.).
Ed. by Carl Murchison. Quarterly. Child behavior, animal behavior, and comparative psychology. Founded 1891.

Psychological Review—Princeton, N. J.: Psychological Review Company.
Subscription $5.50. 540 pages annually.
Bi-monthly. General. Founded 1894. Edited by Herbert S. Langfeld.

Psychological Monographs—Princeton, N. J.: Psychological Review Company.
Subscription $6.00 per vol. 500 pp. Founded 1895. Edited by Joseph Peterson.
Published without fixed dates, each issue one or more researches.

Psychological Index—Princeton, N. J.: Psychological Review Company.
Subscription $4.00. 400-500 pp. Founded 1895. Edited by W. S. Hunter and R. R. Willoughby. An annual bibliography of psychological literature.

Psychological Bulletin—Princeton, N. J.: Psychological Review Company.
Subscription $6.00. 720 pages annually. Psychological literature.
Monthly (10 numbers). Founded 1904. Edited by Edward S. Robinson.

Archives of Psychology—Columbia University P. O., New York City.
Subscription $6. 500 pp. per vol. Founded 1906. Ed. by R. S. Woodworth.
Published without fixed dates, each number a single experimental study.

Journal of Abnormal and Social Psychology—Eno Hall, Princeton, N. J.
Subscription $5; foreign $5.25. Edited by Henry T. Moore. Quarterly. 448 pages ann. Founded 1906. Abnormal and social.

Psychological Clinic—Philadelphia: Psychological Clinic Press.
Subscription $3.00. 288 pages. Ed. by Lightner Witmer. Founded 1907.
Without fixed dates (quarterly). Orthogenics, psychology, hygiene.

Psychoanalytic Review—Washington, D. C.: 3617 10th St., N. W.
Subscription $6. 500 pages annually. Psychoanalysis.
Quarterly. Founded 1913. Ed. by W. A. White and S. E. Jelliffe.

Journal of Experimental Psychology—Princeton, N. J.:
Psychological Review Company, 900 pages annually. Experimental.
Subscription $7.00. Founded 1916. Bi-monthly. Ed. by S. W. Fernberger.

Journal of Applied Psychology—Indianapolis: C. E. Pauley & Co.
Subscription $5. 500 pages annually. Founded 1917.
Bi-monthly. Edited by James P. Porter, Ohio University, Athens, Ohio.

Journal of Comparative Psychology—Baltimore: Williams & Wilkins Co.
Subscription $5 per volume of 450 pages. Two volumes a year.
Founded 1921.
Bi-monthly. Edited by Knight Dunlap and Robert M. Yerkes.

Comparative Psychology Monographs—Baltimore: Johns Hopkins Press.
Subscription $5. 500 pages per volume. Edited by Knight Dunlap.
Published without fixed dates, each number a single research. Founded 1922.

Genetic Psychology Monographs—Worcester, Mass.: Clark University Press.
Subscription $14.00 per year; $7.00 per vol. 1000 pages ann. (2 vols.).
Ed. by Carl Murchison. Monthly. Each number one complete research. Child behavior, animal behavior, and comparative psychology. Founded 1925.

Psychological Abstracts—Eno Hall, Princeton, N. J. Edited by W. S. Hunter and R. R. Willoughby. Subscription $6.00. Monthly. 700 pages annually. Founded 1927.

Journal of General Psychology—Worcester, Mass.: Clark University Press.
Subscription $14.00 per year; $7.00 per vol. 1000 pages ann. (2 vols.).
Edited by Carl Murchison.
Quarterly. Experimental, theoretical, clinical, and historical psychology. Founded 1927.

Journal of Social Psychology—Worcester, Mass.: Clark University Press.
Subscription $7.00. 600 pages annually. Edited by John Dewey, Carl Murchison, and international co-operating board. Quarterly. Political, racial, and differential psychology. Founded 1930.

INCHES
METRIC
C2399